Essays on Beliefs, Identity and Moral Values

Dissertation
submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wirtschaftswissenschaften, Dr. oec.
(corresponds to Doctor of Philosophy, PhD)

presented by
Florian Schneider
from Rüti, ZH

approved in April 2020 at the request of
Prof. Dr. Roberto A. Weber
Prof. Dr. Björn Bartling
Prof. Dr. Michel Maréchal

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 01.04.2020

The Chairman of the Doctoral Board: Prof. Dr. Steven Ongena

Acknowledgments

First, I want to thank my supervisor Roberto Weber for his continual support, his guidance and his encouragement. I am grateful to Björn Bartling and Michel Maréchal, my secondary supervisors, for their valuable feedback and Uri Gneezy for inviting me to visit the University of San Diego and for his insightful comments.

I enjoyed working with my coauthors, Fanny Brun, Nadja Ging-Jehli, Shimon Kogan, Martin Schonger, Ivo Schurtenberger, and Roberto Weber.

I also would like to thank the behavioral group at University of Zurich, in particular Chiara Aina, Ernst Fehr, Lea Heursen, Yi-Shan Lee, Eva Ranehill, Julien Senn, Vanessa Valero, Sili Zhang, Florian Zimmermann and Christian Zünd. All of you contributed to my research through numerous discussions and continuous attendance at my presentations.

I thank my office mates in Zurich, Adrian Bosshard, Giacomin Favre, Andreas Haller, Yi-Shan Lee, Ivo Schurtenberger and Sili Zhang, and in San Diego, Binnur Balkan and Pol Campos-Mercade, for many inspiring discussions.

Finally, I want to thank my family, Alexandra Wepfer and Eva, Felix and Kathrin Schneider, for their constant support.

Contents

Chapter 1: Introduction

In my dissertation, I study how beliefs, identity concerns (or, image concerns) and moral values affect decision making and, ultimately, economic outcomes. Within this broader research area, I investigate how heterogeneity in moral values and image concerns impacts individual choices in markets, market prices and the outcomes obtained by different types of individuals, including income and occupation (Chapter 2 and 3). Moreover, I study how image concerns shape beliefs ("motivated beliefs") in strategic settings (Chapter 4).

In Chapter 2 of my thesis, I study how heterogeneity in social-preferences and image concerns affect outcomes in labor markets. This project is joint work with Fanny Brun and Roberto A. Weber. We study labor markets for jobs that are widely perceived as involving immoral acts. Examples for such jobs include marketing of tobacco products, predatory lending and manufacturing weapons. Our work is focused on two specific hypotheses that arise from a simple model of compensating wage differentials. First, we investigate whether the aversion among many individuals to performing jobs generally perceived as immoral contribute to immorality wage premiums, a form of compensating wage differential. Second, we study whether individuals least concerned with having a good self- and social-image select into such jobs.

We find that immoral work commands a wage premium over comparable work that is not immoral. We show this using Swiss labor market data, where we attempt to control for observable worker and industry characteristics and in laboratory labor markets—which vary, by treatment, only whether being employed requires doing something immoral while holding other aspects of the job constant. We also provide evidence of sorting by immoral types into immoral work. We classify subjects into types using two different measures of individuals' aversion to acting immorally, one from a survey and one from a behavioral task. In the laboratory labor markets, the immoral types are hired more often for immoral work. In our survey data, the immoral types report a significantly greater willingness to work in firms and industries that are perceived to be immoral.

These findings are problematic as immoral jobs and industries often have a great potential to do societal harm; social welfare will therefore likely be higher when workers in such industries voluntarily internalize the negative impacts of their actions and forgo potentially profitable opportunities. However, our evidence suggests that it is the least moral

types who will sort into these industries and that, therefore, labor market sorting will make it less likely that such internalization will occur.

Heterogeneity in moral values and image concerns could also play an important role in markets for consumption goods. In Chapter 3, I study whether people use consumption to signal their moral values to others. Models of identity signaling predict that customers signal their desirable characteristics (or, "types") to themselves and others by avoiding products popular among people with undesirable characteristics and by conforming to product choices of people with desirable characteristics.

Suppose, for example, that the customer pool of a product largely consists of consumers with certain moral values. This resembles the case of the clothing brand Lonsdale in the 1990s; it was public knowledge in Germany that neo-Nazis made up a large share of Lonsdale's customers. If an individual consumes the product in public, others might confuse her with the typical consumer of the product, and attribute the typical consumer's type to her. As a result, many consumers might avoid the product.

In Chapter 3 of my thesis, I provide evidence that consumers indeed care about the type-composition of products' customer pools. My evidence comes from controlled experimental settings that allow me to manipulate the type-composition of products' customer pools while keeping other aspects of the choice environment constant. In a first study, I investigate consumption in a public setting in which subjects' social-images are at stake. In a second study, I investigate (non-)conformity in a double-blind setting in which subjects' self-images are at stake. In both studies, I find that participants' willingness to pay is substantially lower for a product that is popular among people with undesirable moral values than for a product that is popular among people with desirable moral values.

In Chapter 4 (joint work with Nadja R. Ging-Jehli and Roberto A. Weber), I study motivated beliefs in strategic settings where individuals have to form beliefs about the likely behavior of opponents. Strategic beliefs are typically assumed to be determined by the structure of the game and beliefs about others' preferences or rationality. However, in light of the apparent ease with which people bias their beliefs in self-serving ways in other contexts, it seems plausible that they may similarly bias their beliefs about others' actions when doing so can justify acting in a selfish way that harms others. Indeed, a recent paper by Di Tella, Perez-Truglia, Babino and Sigman (2015) provides evidence consistent with the idea that people engage in such "strategic cynicism." Specifically, they demonstrate that people with a greater opportunity to take from another person believe that this opponent is more likely to act in a greedy and harmful manner.

Our study investigates the phenomenon of strategic self-deception, although we initially approach this question in a different manner from Di Tella, et al. Rather than testing whether people with a *greater* incentive to take from others adopt *relatively* more negative beliefs about these opponents, as they do, our focus is on whether people with the opportunity to take from others adopt beliefs that are biased in comparison to the beliefs of neutral outsiders with no incentive to view others self-servingly. That is, we test the extent to which individuals with an incentive to engage in strategic cynicism adopt beliefs that are negatively biased in *absolute* terms. In contrast, Di Tella, et al., study a *relative* form of this bias, investigating whether one group's beliefs are more negative—or, critically, less positive—than those of another group. In contrast with Di Tella, et al. (2015), we find no evidence that individuals engage in "strategic cynicism."

We reconcile the discrepancy, using Di Tella, et al.'s, data, a simple model of belief manipulation and a novel experiment that replicates and extends Di Tella, et al. Across three datasets, we find no evidence of *negatively* biased beliefs. However, Di Tella, et al.'s, results and our data indicate that those with a greater incentive to view others' intentions cynically exhibit relatively less *positive* beliefs. Thus, to the extent that bias exists in people's beliefs about a counterpart's actions, it appears to be one of positivity rather than cynicism. This positivity bias is in line with another form of motivated belief; namely, in the kinds of interactions we study here, individuals seem motivated to convince themselves of the deservingness of the counterpart, and end up with beliefs that are often too positive. The finding that people are too positive about other players' kindness supports a general tendency for distorted beliefs to lie in the direction of positivity and optimism rather than the opposite.

Chapter 2: Sorting and wage premiums in immoral work

Joint with Fanny Brun and Roberto A. Weber

Abstract

We use surveys, laboratory experiments and administrative labor-market data to study how heterogeneity in the perceived immorality of work and in workers' concerns with acting immorally interact to impact labor market outcomes. Specifically, we investigate whether individuals least concerned with acting morally select into jobs generally perceived as immoral and whether the aversion among many individuals to performing such acts contributes to immorality wage premiums, a form of compensating differential. We obtain two measures of an individual's aversion to performing immoral acts, one from a behavioral laboratory task and the other from survey items. These two measures predict laboratory labor market outcomes and expected outcomes in non-laboratory labor markets for "immoral" work. In the laboratory, immoral types are more likely to be employed and obtain higher wages when a job requires performing immoral acts. In our survey data, immoral types express a greater willingness to work in firms and industries rated by others as immoral. We also document that wages are higher in such immoral industries, both in laboratory and non-laboratory labor markets.

Citation

Schneider, F. H., Brun, F. and Weber, R. A. (2020) "Sorting and wage premiums in immoral work," working paper.

2.1 Introduction

Immoral behavior in firms has the potential to cause significant social harm. For example, scandals in the financial industry involving the intentional sale of toxic assets to unsuspecting clients (US Department of Justice, 2016) and the aiding of tax evasion (Hill, 2012) create significant burdens for public funds and for trust in the financial sector. Tobacco companies have long been accused of regularly engaging in unethical marketing tactics such as misleading the public about the harmful effects of smoking (Heath, 2016) and developing marketing strategies to attract underage smokers (Bates and Rowell, 1998). Aggressive marketing of prescription opioids by pharmaceutical firms is responsible for a serious public health crisis (Okie, 2010; Case and Deaton, 2015). In cases like these, as well as in many less extreme examples, corporate activities that many regard as "immoral" or "unethical"—but are nevertheless profitable—may have serious negative impacts on social welfare.

Rather than representing isolated incidents, there exists a widespread impression that some jobs—e.g., marketing tobacco products, manufacturing weapons—likely involve inherently immoral acts. Conventional wisdom further posits that such jobs disproportionately attract those individuals who experience the least displeasure from acting immorally and that workers performing these kinds of jobs receive high wages for their unethical conduct—a form of compensating differential driven by the aversion to performing immoral acts.[1] Thus, "immoral" work shares features with other aspects of employment that people may find heterogeneously aversive, such as risk of physical harm (Rosen, 1986).

However, despite the intuitive appeal of such a connection, there is little empirical evidence that links heterogeneity in the willingness to perform work that most people perceive as immoral to resulting differential labor market outcomes. Moreover, there are many reasons to believe that such concerns may be mitigated in competitive markets (Levitt and List, 2008), where wages are set by the moral concerns of the marginal worker and where repeatedly forgoing profitable job opportunities may lessen workers' concerns for avoiding immoral behavior.

In this paper, we provide novel evidence testing the above relationships between the perceived immorality of work, workers' heterogeneous concerns for morality and outcomes

[1] This view dates back to Adams Smith (1776; Book I; Ch. X), who wrote that "The exorbitant rewards of players, opera-singers, opera-dancers, etc., are founded upon [..] the discredit of employing them in this manner. [..] Should the public opinion or prejudice ever alter with regard to such occupations, their pecuniary recompense would quickly diminish. More people would apply to them, and the competition would quickly reduce the price of their labour." At the time, such professions were seen as morally tainted; Smith equated them with "a sort of public prostitution."

in labor markets. We do so with a combination of surveys, laboratory experiments and administrative labor market data, in which we obtain measures of individuals' concerns for morality and relate these to variation in the morality of work. The administrative data provides the clearest evidence of the economic relevance of these relationships, but in these data "immoral" industries might differ in many aspects from other industries (for example, in regulation threats and litigation risks), meaning that it is impossible to establish causality. The control provided by laboratory experiments, however, allows us to observe the kinds of outcomes that arise as the nature of work changes *only* in the extent to which it is immoral. We additionally use complementary evidence from different surveys to obtain insights into relationships between individuals' concerns for morality and the morality of different firms and industries.

Our work is focused on two specific hypotheses that arise from a simple theoretical analysis of how individuals' heterogeneous aversion to performing immoral work may interact with jobs that vary in the immorality of the work they require. We do not attempt to provide a novel theoretical contribution, but rather use simple economic analysis as a framework for formalizing standard intuitions and guiding our empirical research. The two hypotheses reflect the widely held perceptions that we note above: first, that more immoral work should yield higher wages—as long as workers care enough about morality—and, second, that immoral work should attract those workers who care the least about morality. We then use laboratory experiments, survey evidence and administrative data to test these hypotheses. A critical novel contribution of our work is to directly relate heterogeneity in individual concerns for morality to differences in labor market outcomes.

Table 1. Overview of our evidence on wage premiums and sorting

	Laboratory labor market	*Labor markets outside the laboratory*
Immorality premium	Causal evidence for a wage premium for immoral work (Section 4; Figure 5)	Correlation between perceived industry immorality and wages in Swiss Labor Force Survey (Section 3; Figure 1, Table 2)
Sorting	Immoral types are more likely to be hired, but only for immoral work (Section 4; Figure 6, Table 5)	Immoral types state a greater willingness to work in firms and industries perceived to be immoral (Section 5; Figure 9, Table 7)

Our results provide clear support for the above two hypotheses both in and out of the laboratory. Table 1 gives an overview of our main findings and refers to the key figures and tables.

First, we show that immoral work commands a wage premium over comparable work that is not perceived as immoral. We show this in Section 3 using administrative labor market data, where we find that industries that are perceived to be immoral yield higher average wages, controlling for observable worker and industry characteristics. Moreover, in our laboratory labor markets—which vary, by treatment, only whether being employed requires doing something immoral while holding other aspects of the job constant—we observe a causal relationship indicating that wages are persistently higher for immoral work (Section 4). This wage premium is large and does not decrease with market experience, reflecting a strong and stable aversion to immoral work on the part of our laboratory participants.

Second, we provide evidence of sorting by immoral types into immoral work, both in the laboratory and in the field. We obtain two measures of individuals' aversion to acting immorally, one from a behavioral task and one from a survey, and find that these two measures are positively correlated. These measures of participants' immorality predict individual labor market outcomes: in our laboratory labor markets (Section 4), immoral types are employed more frequently, but only when work involves doing something immoral—i.e., there is no difference when the labor market does not involve immoral work. In our survey data (Section 5), immoral types report a significantly greater willingness to work in firms and industries that others perceive to be immoral.[2] Aside from confirming one of our main hypotheses, this finding also indicates that the perceived immorality of industries and firms might be identifiable by a revealed (or stated) preference approach—that is, the immorality of firms or industries may be identifiable by the degree to which they are relatively more attractive employment opportunities for those individuals less concerned with morality.

Our work provides the first evidence documenting a differential willingness of heterogeneous moral types to work in jobs and industries that vary in their perceived immorality. Importantly, this sorting persists with experience in our laboratory experiment and extends to stated preferences regarding real labor-market outcomes. We also connect the heterogeneity in the immorality of work to wage premiums. In our field data this connection

[2] While these job choices are hypothetical, Wiswall and Zafar (2018) provide evidence that such stated preferences predict subsequent employment.

is correlational and thus subject to caution in drawing interpretations from the relationship, but our laboratory study demonstrates a causal relationship.

The closest prior evidence supporting the relationships we investigate comes from studies documenting positive correlations between the perceived immorality of work and the wages obtained by workers in those "immoral" firms or industries (Frank, 1996; Moffatt and Peters, 2004; Arunachalam and Shah, 2008; Edlund, Engelberg and Parsons, 2009). This is similar to the correlational evidence we provide in Section 3, but prior evidence focuses on more limited samples. For example, Frank (1996) used data from a Cornell University employment survey that included graduates' occupations, reported salaries and employers.[3] He then asked students in a business ethics class to rate the "social responsibility" of the most common occupations and employers of Cornell graduates. A regression controlling for other observable characteristics—such as a student's major, grades and gender—reveals a premium for occupations and employers that are rated as socially irresponsible.

Frank's evidence is consistent with the notion that concerns for avoiding immoral work produce differential wages across occupations and industries. However, there remain important gaps in documenting that the relationships observed by Frank are really the result of sorting, heterogeneous moral preferences and compensating differentials. Most obviously, correlational evidence between wages and the perceived immorality of work might result from other unobserved characteristics of workers and the work activities. For instance, Moffatt and Peters (2004) document a wage premium for prostitution, which they attribute to a compensating differential (see, also, Arunachalam and Shah, 2008; Edlund, Engelberg and Parsons, 2009); but it is unclear whether the compensation is for the perceived immorality of the work or other aversive aspects of the job (Edlund and Korn, 2002; Gertler, Shah and Bertozzi, 2005).[4] Moreover, such studies fail to measure a critical element—the identification of workers' heterogeneous concerns for morality—as a key driver of the relationship. While there is some correlational evidence that people in some industries and professions exhibit lower concerns for morality (Carter and Irons, 1991; Sjöberg and Engelberg, 2009; Dur and Zoutenbier, 2014),[5] no study identifies that these concerns for

[3] Note that these wages may not correspond to (average) industry wages because his sample is very selective. Unlike Frank (1996), we use a sample that is representative of the national work force in Switzerland.

[4] Related work in finance (e.g., Fabozzi, Ma and Oliphant, 2008; Hong and Kacperczyk, 2009; Colonnello, Curatola and Gioffré, 2019) demonstrates that investing in firms that engage in immoral activities ("sin stocks") yields higher returns. However, other industry and firm characteritics, such as litigation risk, may also differ between these types of investments (Blitz and Fabozzi, 2017).

[5] Gregg, et al. (2011) find that non-profit employees are more likely to do unpaid overtime. In line with selection, they find that individuals do not adjust their behavior when they change sectors. Note, however, that this finding could also be explained by a persistent effect of company culture. Fisman et al. (2015) find some

morality drive differential selection into different kinds of work rather than the relationship being perhaps the other way around (Frank, Gilovich and Regan, 1993; Cohn, Fehr and Maréchal, 2014).

We also contribute to the literature on morality and markets (e.g., Falk and Szech, 2013; Bartling, Weber and Yao, 2015; Kirchler, Huber, Stefan and Sutter, 2016). Levitt and List (2007, 2008) question whether social preferences matter in markets, due to factors including high stakes, market competition and experience. Our estimates from the labor market data indicate a substantial wage premium for immoral work: individuals working in industries perceived as highly immoral are estimated to have 35 percent higher wages than people working in morally neutral industries. Our survey data indicates that social preferences predict sorting into industries. Finally, in the laboratory labor market, we find that neither the wage premium nor sorting diminish with market experience. These findings suggest that social preferences matter for markets, and that they can impact both market wages and individual market outcomes.

Finally, we also contribute to the literature on compensating differentials. A substantial number of studies investigate whether nonprofit employees earn less than for-profit employees (e.g., Leete, 2001; Mocan and Tekin, 2003; Ruhm and Borkoski, 2003; Benedict, McClough and McClough, 2006; Jones, 2015). These studies yield mixed correlational evidence, likely due to methodological challenges in estimating compensating wage differentials using observational data (see the discussion in Mas and Pallais, 2017). Recent papers on compensating differentials rely on experimental methods and/or stated preferences instead (Eriksson and Kristensen, 2014; Pörtner, Hassairi and Toomim, 2015; Carpenter, Matthews and Robbett, 2017; Mas and Pallais, 2017; Maestas et al., 2018; Wiswall and Zafar, 2018). We differ from this work in that we explore the immorality of work and the aversion to immoral acts as the driving sources of heterogeneity and we present causal evidence on compensating wage differentials in the domain of morality.[6]

---

evidence that distributional preferences (equality-efficiency tradeoffs) of Yale Law School students predict students' career choices. Hanna and Wang (2017) investigate selection into public services in India. In line with their theory that there are more opportunities for corruption in the public sector, they find that students who cheat in a laboratory experiment are more willing to work for the government.

[6] Our study also relates to research on effort and sorting across different kinds of work by "mission-oriented" types (Besley and Ghatak, 2005; Prendergast, 2007; Delfgaauw and Dur, 2008; Ariely, Bracha and Meier, 2009; Dal Bó, Finan and Rossi, 2013; Fehrler and Kosfeld, 2014; Tonin and Vlassopoulos, 2015; Carpenter and Gong, 2016; Cassar and Meier, 2018; Cassar, 2019; Deseranno, 2019; Dur and van Lent, 2019), though this line of research typically focuses on worker motivation and effort and not on morality and labor demand. Moreover, our findings also loosely relate to recent studies on how income relates to moral behavior (Bartling, Valero and Weber, 2018; Andreoni, Nikiforakis and Stoop, 2017).

The rest of our paper proceeds as follows. The next section presents a simple theoretical framework that we use to motivate the relationships that one would expect to see in labor markets under heterogeneity in moral concerns and work characteristics. We then use Swiss labor market data to investigate the relationship between the immorality of work and wages. Section 4 presents the design and results of our laboratory experiment. Section 5 gives the design and results of our survey study.

We conclude in Section 6 by discussing several implications of our findings for policy. For example, sorting by individuals more willing to engage in immoral acts may exacerbate the potential social harm produced by industries and firms with production technologies that involve negative externalities. That is, from society's perspective we may want those people most concerned with acting morally working in industries with the greatest potential for producing harm, but our results suggest the opposite may happen. Such resulting impacts are of critical importance, since the design of policies and market features aimed at mitigating negative social impacts need to take into account heterogeneity in the preferences of individuals ultimately making decisions.

2.2 A simple model of heterogeneous moral concern in labor markets

In this section, we introduce a simple stylized model of labor markets with varying degrees of perceived job immorality and heterogeneity in concern for moral behavior among workers. We use the theoretical results to guide our investigation of immoral labor markets.

We examine a single labor market for a job, $j \in J$. The job might involve doing immoral work. Firms decide whether to hire a worker to do j at the market wage, w, and workers decide whether to accept work j for the market wage. Workers differ in their concerns for morality. We then investigate how the equilibrium wage and selection in this labor market change when we increase the immorality of j. Our framework is a simplification of the theoretical literature on compensating wage differentials (see, e.g., Rosen, 1986).[7] We do not seek to expand this literature, but rather to apply it to a context in which the relevant job dimension is immorality.

[7] Unlike most models of compensating wage differentials, we do not have multiple labor markets, but only one, along with a fixed outside option. In our laboratory experiment, we also assign subjects to one labor market. This abstraction simplifies both the theory and the experiment. However, we show in Appendix E that our model allows for an interpretation with two jobs, an immoral job and a neutral job. Our results also apply to such a context.

The immorality of j is measured by a function $I: J \to [0, \infty)$, where $I(j') > I(j)$ means that job j' is more immoral than job j, and $I(j) = 0$ means that the job j involves no immoral acts. The set of immoral jobs is $J^{IM} = \{j \in J: I(j) > 0\}$.

Labor demand is represented by an interval of firms, $k \in [0,1]$. Firms' behavior is given by the labor demand function, $D: \mathbb{R} \times J \to [0,1]$, with $lim_{w \to \infty} D(w,j) = 0$, $D(w,j) = 1$ for $w \leq 0$, D continuous in w and D strictly decreasing in w on $[0, \infty)$. In addition, we assume that an increase in the immorality of the job does not decrease profitability of labor, that is, $I(j') > I(j)$ implies $D(w,j') \geq D(w,j)$ for all w.[8]

Labor supply consists of an interval of workers, $i \in [0,1]$. Each worker has reservation utility $\underline{u} \geq 0$, and the utility of accepting job j of a worker of type i is given by:[9]

$$u_i^{accept}(j,w) = w - c - \theta_i * I(j),$$

where $c \geq 0$ is the worker's cost of effort, which is independent of j. The parameter $\theta_i \geq 0$ measures how much the worker cares about the immorality of the job and is distributed according to a cumulative density function $F \in \mathcal{F}_\theta$. The set $\mathcal{F}_\theta$ consists of all density functions F that are continuous, strictly increasing on $[0, \infty)$, and with $F(0) = 0$.[10] The indirect utility of a worker of type i is then given by $v_i(j,w) = max\{\underline{u}, u_i^{accept}(j,w)\}$. Workers' behavior determines the labor supply, $S: \mathbb{R} \times J \to [0,1]$. If $j \in J^{IM}$, every worker with $\theta_i \leq \frac{w - \underline{u} - c}{I(j)}$ accepts the job. Labor supply is therefore $S(w,j) = F(\frac{w - \underline{u} - c}{I(j)})$.[11]

Using this framework, we can now consider the equilibrium properties of this type of market. The equilibrium wage, $w^*(j)$, is implicitly defined by $S(w^*(j), j) - D(w^*(j), j) = 0$.[12] The following Lemma states that for every $j \in J$, $w^*(j)$ exists and is unique (all proofs are in Appendix D).

---

[8] In our experiment, we vary the immorality of the job, but fix labor demand, that is, $D(w,j') = D(w,j)$ for all w and all $j, j' \in J$. If an increase in immorality were to decrease profitability, there would be no incentives for firms to operate in immoral industries. Heidhues, Kőszegi and Murooka (2017) provide a basis for why deceptively marketed socially harmful products may be more profitable in the presence of naïve consumers. In Appendix E, we provide a behavioral foundation for the labor demand.

[9] Models about "mission-oriented" employees commonly assume very similar, additive utility functions (e.g. Cassar and Meier, 2018), with the main difference that $-I(j) * \theta_i$ is replaced by a positive term, the "meaningfulness of work" multiplied by how much the individual cares about meaning.

[10] Note that $F(0) = 0$ implies that no worker likes to do immoral jobs (Rosen, 1986, p. 645, makes a similar assumption).

[11] The assumptions on F (together with the properties of a cdf) imply that S is continuous and strictly increasing in w on $[\underline{u} + c, \infty)$, $lim_{w \to \infty} S(w,j) = 1$, and $S(w,j) = 0$ for all $w \leq \underline{u} + c$.

[12] Note that for $j \in J \setminus J^{IM}$, S is a correspondence. For this case, $w^*(j)$ is defined by $D(w^*(j), j) \in S(w^*(j), j)$. Moreover, $w^*(j)$ depends on F. When necessary (Proposition 4) we will make this explicit by writing $w^*(j, F)$ instead of $w^*(j)$.

Lemma. *For all* $j \in J^{IM}$, $w^*(j)$ *exists, is unique and is in* $(\underline{u} + c, \infty)$. *For all* $j \in J \setminus J^{IM}$, $w^*(j) = \underline{u} + c$.

In the following, we derive four properties of labor markets with immoral jobs. While straightforward, we use these results to make predictions for our empirical work. In particular, the first two propositions derive the primary hypotheses that we test across all of our analysis.

Proposition 1 shows that there is an immorality premium for immoral jobs: an increase in the immorality of a job decreases supply and therefore increases the equilibrium wage.

Proposition 1. (Immorality premium) *For all* $j, j' \in J$ *with* $I(j) < I(j')$, $w^*(j) < w^*(j')$.

The following Corollary further shows that this wage premium will be insignificant if workers do not sufficiently care about morality.[13]

Corollary. *For all* $j, j' \in J$ *with* $I(j) < I(j')$ *and* $\varepsilon > 0$, *there exists* $G \in \mathcal{F}_\theta$ *such that* $w^*(j', G) - w^*(j, G) \leq \varepsilon$.

Formally, the Corollary shows that there are distributions of moral types such that the wage differentials are arbitrary small ($\leq \varepsilon$).

Second, the immoral types, or to be precise, the types that care least about the immorality of a job ($\theta_i \leq \frac{w^*(j)-\underline{u}-c}{I(j)}$), sort into accepting immoral jobs, while those workers more concerned with morality ($\theta_i > \frac{w^*(j)-\underline{u}-c}{I(j)}$), refuse to do the job for the equilibrium wage.[14] This is formally shown in Proposition 2.[15]

Proposition 2. (Sorting) *For all* $j \in J^{IM}$, *worker* i *is hired iff* $\theta_i \leq \frac{w^*(j)-\underline{u}-c}{I(j)} \equiv \underline{\theta}(j) > 0$.

[13] Becker (1957) made a very similar point in his analysis of discrimination: discrimination will only affect wages if there is sufficiently large share of discriminating employers.

[14] Note that this perfect sorting according to θ is an extreme, and admittedly unrealistic, case. Heterogeneity in the costs of effort, reservation utility or productivity implies partial sorting according to θ. We do not incorporate more than one dimension of heterogeneity in our model to keep it simple. Heterogeneity in both productivity (earnings capacity) and risk preferences is investigated in Garen (1988) and Hwang, Reed and Hubbard (1992).

[15] Note that for any $a \in \mathbb{R}_{>0}$, $F(a) > 0$ because $F(0) = 0$ and F is strictly increasing on $[0, \infty)$. Therefore $F(\underline{\theta}(j)) > 0$, implying that some workers are hired. This is also important for most of the other propositions.

Proposition 2 is critical to the notion that immorality wage premiums are driven by those who find immoral work most distasteful opting out of such jobs. This important relationship has, to our knowledge, not been previously empirically tested.

Our next two propositions are less central to our purposes, but nevertheless provide some useful and testable insights into behavior in immoral labor markets. Proposition 3 shows that immoral types profit from an increase in the immorality of work.

Proposition 3. For all $j, j' \in J$ with $I(j) < I(j')$, there exists $\tilde{\theta}(j, j') > 0$ such that $v_i(j', w^*(j')) > v_i(j, w^*(j))$ iff $\theta_i < \tilde{\theta}(j, j')$.

More precisely, there are always some types who are sufficiently unconcerned with morality who are hired in an immoral market and are overcompensated by the immorality premium.

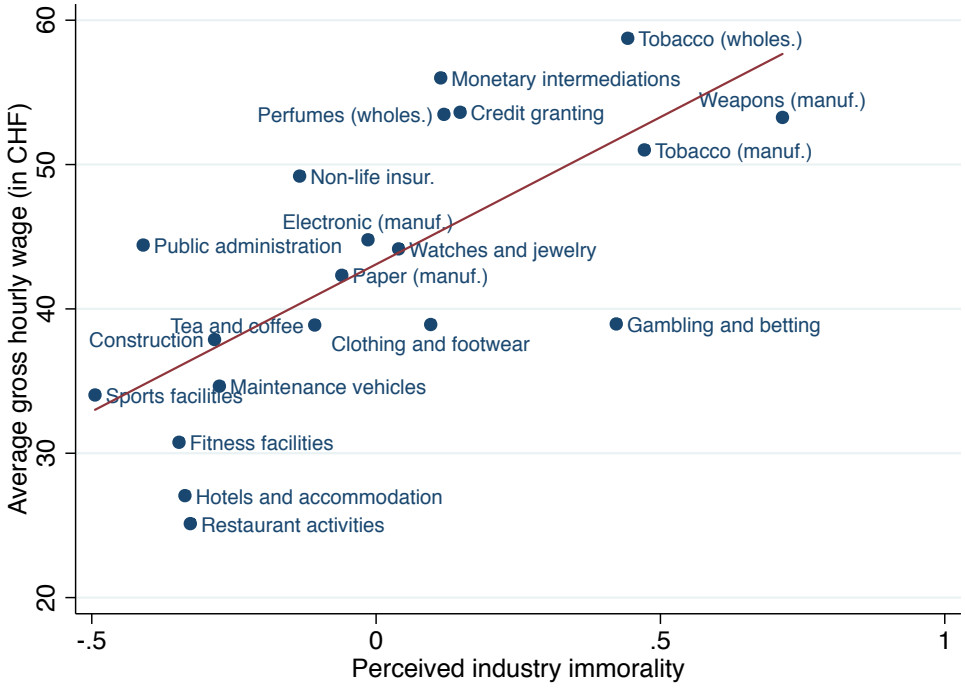
Finally, Proposition 4 shows that moral behavior (refusing to do the immoral job) can have positive externalities for the most immoral types. If the distribution of types shifts toward more concern for morality—in the sense of stochastic dominance—then the supply at any given wage decreases, thereby increasing the equilibrium wage and the utility of those least concerned with morality. For instance, any completely immoral types (i.e., $\theta_i = 0$) will always benefit from a higher wage produced by increased moral concerns.

Proposition 4. *For all* $j \in J^{IM}$ *and* $F, G \in \mathcal{F}_\theta$ *with* $F(x) < G(x)$ *for all* $x > 0$*, there exists* $\hat{\theta}(j, F) > 0$ *such that* $v_i(j, w^*(j, F)) > v_i(j, w^*(j, G))$ *iff* $\theta_i < \hat{\theta}(j, F)$.

2.3 Evidence of an immorality premium from the Swiss Labor Force Survey

To begin our analysis, we explore whether individuals working in industries generally perceived as immoral receive an immorality premium. We use data from the Swiss Labor Force Survey (SLSF)—a representative sample of the Swiss labor force compiled by the Swiss Federal Statistical Office—to explore the relationship between the perceived immorality of work and the portion of wages that cannot be explained by observable worker and industry characteristics.

Figure 1. Correlation between wages and perceived industry immorality



Source: Weighted data from the SLFS, years 2010-2016 (wage) and our own survey (perceived industry immorality). Notes: Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral. Real gross hourly wage in 2010 CHF. N = 32,638.

We first identified industries that we jointly perceived as likely involving work activities widely seen to be immoral; we did so before looking at any data, including wages, from these industries.[16] This yielded six "immoral" industries: gambling and betting activities, monetary intermediations, credit granting, manufacture of tobacco products, wholesale of tobacco products and manufacture of weapons and ammunitions. We then chose comparison industries from within the same industrial branch with similar distributions of education levels, as well as nine additional industries representing large shares of employment in Switzerland.[17] We did not look at wages when selecting these industries.[18]

[16] Specifically, we started with the complete list of industries listed in the SLFS dataset. Each of the three authors went through the list and indicated any industries that he or she perceived as having (or, being widely perceived to have) a significant immoral component. We selected those industries for which all three authors agreed. We proceed this way, rather than using the entire set of Swiss industries, to keep the number of questions we ask in subsequent surveys manageable.

[17] We chose five comparison industries: non-life insurance (for monetary intermediations; credit granting), organization and operation of sport facilities (for gambling and betting activities), processing of tea and coffee (for manufacture of tobacco products), manufacture of electronic components (for manufacture of weapons and ammunitions), wholesale of perfume and cosmetics (for wholesale of tobacco products).

[18] One might be concerned that we choose the sample of industries ourselves. In Appendix B we provide evidence for an immorality premium that does not rely on our sample of industries. First, we show that "immoral industries" are also estimated to pay a wage premium when we add *all* other industries as control

We next obtained independent ratings of the perceived morality of the selected industries, by asking a sample of 177 students on the campus of the University of Zurich and the ETH to rate each industry on a 5-point Likert scale ranging from "very moral" to "very immoral" and re-scaled the responses to lie on the -1 to 1 interval. (These survey data were collected as part of our survey studies, which we describe in more detail in Section 5.) We interpret this variable as a noisy measure of the immorality of working in industry j, or $I(j)$, a key component of our theoretical model. The mean ratings for each industry are shown on the horizontal axis of Figure 1. They confirm that our initial judgments with respect to the perceived immorality of certain industries are shared by the student sample.

The vertical axis of Figure 1 plots the mean real gross hourly wage (in 2010 Swiss Francs) in each industry. These data are the reported hourly wages of employees surveyed as part of a national representative panel. We use data from the 2010 to 2016 waves. The strong positive relationship supports the hypothesis that work in less moral industries yields a wage premium.

Of course, the relationship in Figure 1 ignores the potential role of individual worker characteristics, which may vary across industries, and other characteristics of the industries themselves that may partly explain the wage gap. To partially address this concern, Table 2 reports regressions of the hourly wage reported by individuals in different industries on the perceived immorality of each industry, along with several additional control variables. Model 1 displays the results of a simple regression of the natural logarithm of real gross hourly wages on the perceived industry's immorality, supporting the positive relationship in Figure 1. Model 2 adds observable worker and industry characteristics, while Model 3 additionally includes indicator variables for each year and indicator variables for the region where the employer is located.[19] While the addition of these controls lowers the magnitude of the industry immorality coefficient, the immorality premium remains large and statistically significant: according to Model 3, individuals working in an industry as immoral as manufacture of tobacco products (i.e., Perceived immorality = 0.47) have (geometric) mean

industries. Second, we replicate the entire analysis with another set of industries that were selected by research assistants who were not familiar with the research question.

[19] First, note that the number of clusters is relatively small. We also computed the significance levels using the wild bootstrap procedure described in Cameron, Gelbach and Miller (2008) with 400 replications: p-values for perceived industry immorality are 0.070 for Model 1, 0.000 for Model 2, and 0.005 for Model 3. Second, note that numbers of observations differ substantially between industries. If we weight observations by industry size (instead of using survey weights), estimates for perceived immorality are smaller (Model 1: 0.489, Model 2: 0.349, Model 3: 0.344), but still significant (t=3.88, 3.10, and 4.03, respectively). Summary statistics of all variables used in the regressions and all of the industries are provided in Appendix Table A1.

Table 2. Relationship between wages and perceived industry immorality

Dependent variable: ln of real gross hourly wage (in 2010 CHF)			
	(1)	(2)	(3)
Perceived industry immorality	0.929***	0.736***	0.638***
	(2.95)	(4.15)	(3.85)
Age		0.005***	0.005***
		(3.32)	(3.64)
Male		0.204***	0.203***
		(6.60)	(6.80)
Married		0.033*	0.037**
		(1.88)	(2.13)
Education high		0.469***	0.442***
		(8.38)	(8.64)
Education middle		0 .177***	0.170***
		(4.53)	(4.77)
Swiss		0.028	0.036*
		(1.15)	(1.97)
Experience		0.005**	0.005***
		(2.81)	(3.17)
Full-time equivalent		-0.037	-0.038
		(-0.51)	(-0.54)
Managerial duties		0.059	0.065
		(0.71)	(0.82)
Industry sales		0.027	0.034
		(0.94)	(1.25)
Industry size (employees)		0.001***	0.001***
		(3.77)	(3.69)
Constant	3.759***	2.930***	
	(70.71)	(26.43)	
N	32,638	32,638	32,638
Adjusted R^2	0.140	0.379	0.397
Year FE	No	No	Yes
Region FE	No	No	Yes

Source: Weighed data from the SLFS, years 2010-2016 (wage and demographics), STATENT, years 2011-2016 (industry size, industry sales), Value Added Tax Statistics, years 2010-2016 (industry sales) and our own survey (perceived industry immorality).

*Notes: Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral. Control variables: Male in {0, 1}, Married in {0, 1}, Education high: higher vocational education and training or university/college, Education middle: apprenticeship, full-time vocational school, matura or pedagogical training, Education low (reference category): compulsory schooling or pre-vocational education, Swiss in {0, 1}, Experience = number of years in the firm, Full-time equivalent = (working hours /42), set to 1 for working hours >= 42, managerial duties in {0, 1}, Industry size = number employees in this industry / 1000 (2010 data is not available, we substitute it with 2011 data), Industry sales = Industry sales/number employees in this industry. Model (3) controls for company region fixed effects (26 Swiss cantons) and year fixed effects (2010-2016). Standard errors clustered at the industry level, t-statistics in parentheses; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.*

hourly earnings approximately 35 percent higher than people working in a neutral industry (i.e., Perceived immorality = 0).[20]

The above analysis is consistent with the notion that workers are sufficiently concerned about morality such that they are compensated for the perceived immoral nature of some kinds of work (Proposition 1).[21] However, the correlational aspect of the relationship leaves open the question of whether additional unobserved characteristics of the industries may explain the relationship in Figure 1. Moreover, the above analysis does not tell us whether the workers employed in these industries differ in their concerns for morality. In the following, we explore the predictions of our model more carefully—controlling for unobservable aspects of the work and making a clearer connection to subjects' heterogeneous concerns for morality—by means of a laboratory experiment and a survey study. We measure heterogeneity in concerns for morality (θ_i) and relate this heterogeneity to workers' outcomes as labor markets vary in the immoral nature of work ($I(j)$).

2.4 Sorting and wage premiums in a laboratory labor market

The experimental study uses two different subject samples: *workers* and *clients*. Our main focus is on the sample of workers, who participate in our laboratory sessions. These subjects initially complete an on-line questionnaire, and then participate in a laboratory experiment approximately one week later. The questionnaire measured subjects' concern for morality using survey instruments, as well as their preferences regarding future employment possibilities. We discuss the on-line questionnaire in detail in the next section, where we also report the analysis of the resulting data. In this section, we focus on the design and results of the laboratory experiment. In the laboratory, we investigate the causal impact of the immorality of work on labor market outcomes. We do so by exogenously varying the degree of immorality of work while keeping everything else constant. In addition, we measure subjects' concerns for morality and relate them to outcomes in the laboratory labor markets.

In the laboratory sessions, subjects perform two choice tasks. First, we elicit a measure for concerns for morality (corresponding to θ in our model) using an incentivized

[20] We obtain this number by doing the following calculation: $e^{0.638*0.47} - 1 \approx 0.350$.

[21] This finding is in line with reports from companies that had recent scandals or bad moral reputations: Facebook, in wake of the Cambridge Analytica scandal, struggled to attract top talent (CNBC, 2019). Tobacco companies deem difficulties in recruitment arising from their image as an important enough risk to warrant disclosure to regulators and shareholders (British American Tobacco, 2015, p. 37; Philip Morris International Inc., 2015, p. 14).

behavioral task, adapted from Gneezy, Rockenbach and Serra-Garcia (2013), that creates a tradeoff between personal monetary gain and moral conduct. We will use this measure to investigate whether participants with low concerns for morality sort into immoral labor markets. Subjects then participate in a laboratory labor market for 15 periods, in which they submit reservation wages for performing a task. Labor demand is simulated according to a fixed demand schedule, that is, computerized employers automatically submit wage offers.

The key feature of our laboratory experiment is that we vary by treatment only the degree to which work is immoral, while holding constant all other job characteristics, including the specific actions subjects take when employed. We attempt to design an "immoral" act that is unambiguously harmful and for which there is likely widespread agreement regarding its immorality. We opt for an act akin to giving bad financial advice to a non-profit organization, like UNICEF, thereby harming the non-profit's financial standing and thus harming the organization's employees and its ability to help aid recipients, and destroying potential value created by donors' contributions. Providing harmful information and misleading customers is a realistic feature of many existing jobs perceived as immoral.

We operationalize this kind of scenario in our experiment by informing subjects that each session is endowed with an initial donation to a UNICEF fund that provides malaria treatments for children (the aid recipients). These initial donations are linked to a donation generated by a third party's blood donation (the donors).[22] However, the actual final donation for a session is influenced by the behavior of participants in the session. Specifically, subjects in our experiment are hired to provide written advice to a "client" (a subject who participates later and serves a role analogous to the non-profit's employee). We vary, by treatment, whether workers are assigned to a market with neutral jobs that involve honest advice that has little impact on the client and the UNICEF fund or to a market with immoral jobs that involve dishonest advice that hurts the client and UNICEF. A worker's choice of whether to accept work is visible to other workers in the labor market.[23]

We recruit a separate sample of individuals, the *clients*, at public locations. These participants serve two functions. First, they serve in the role of "clients" who receive written recommendations from laboratory subjects and act upon this advice. From these choices, the

[22] Prior to the laboratory session, we approached individuals who had just donated their blood as part of a donation campaign. We asked them whether they would agree that the University of Zurich potentially makes a donation to UNICEF as a complement to their blood donation. Most donors we approached agreed.

[23] This means that our behavioral measure of "immorality" confounds both internally-driven concerns for acting immorally and concerns for being perceived as willing to act immoral by others. Since real-world labor markets typically also confound both motives, we do not draw a distinction in our study, but instead combine both motives to strengthen the (perceived) immoral nature of work.

clients accumulate money and determine the size of a donation to UNICEF. Second, these participants complete a survey in which they evaluate the extent to which various industries and firms are "moral" or "immoral." These are the ratings that we already used in Figure 1 and Table 2.

2.4.1 The laboratory experiment

In the laboratory experiment, sessions consisted of 24 participants. During the experiment, subjects accumulated earnings in "points," which were converted to money at the rate of 20 points = 1 CHF ≈ 1 US Dollar.

Before subjects entered the lab, we took a portrait photograph of each subject to make labor market outcomes public. Participants were asked to make a neutral face while the picture was taken. Next, all participants entered the laboratory. As subjects have the ability to influence the amount of a donation to a UNICEF fund that provides treatment to children with malaria, participants read an information sheet about the consequences of malaria and the need for treatments at the beginning of the study—we adopted wording from UNICEF's public materials and referred to each donation unit as helping to "save a child" from malaria by providing a treatment.

In the following, we describe each of the choices subjects completed in the laboratory session, in detail. We also provide details on the recruitment and role of the clients.

1.4.1.1 Behavioral measure of concern for morality (θ^{Exp})

Participants first played an incentivized game that measures their willingness to lie for personal gain while causing harm to others in a non-market environment. The task builds on a game by Gneezy et al. (2013), and modifies it such that it mimics the consequences of a lie in the immoral treatment in our experimental labor market. Remember that we elicit this measure to investigate sorting into immoral labor markets.

In the game, Participant A privately observes a computerized die roll and sends a message reporting the observed number to Participant B. Participant A may claim that the observed number *r* is either "1", "2," "3," "4," "5," or "6," regardless of the actual number. Participant A receives *100 + 20 * r* points, which means that she has an incentive to lie if *r* is less than 6. Participant B then decides whether "to follow" or "not to follow" the message sent by Participant A. If Participant B does not follow the message, he receives 30 points and the donations to UNICEF are unaffected. If he follows the message and Participant A truthfully reported the observed number, Participant B earns 100 points and the initial

donation to UNICEF is increased by an amount corresponding to one additional anti-malarial treatment.[24] However, if Participant B follows the message and Participant A lied, Participant B does not earn any points and the donation to UNICEF is *reduced* by one additional anti-malaria treatment.

Every participant initially plays the role of Participant A. We use the strategy method to elicit Participant A's message for every possible die roll. This allows us to classify all subjects by their strategies in the game. At the end of the experimental session, 5 of the 24 participants in the session have their role changed from Participant A to Participant B. These Participants B are then matched with five of the remaining Participants A and decide whether or not to follow the corresponding message. All participants whose role is not switched—who remain as Participant A—are paid based solely on their own choice as Participant A, independently of whether or not they are matched with a Participant B.[25]

Participants were informed that, at the very end of the session and after all choices had been made, their decisions as Participant A would be publicly displayed to other participants in the session, along with their portrait photograph. This was all explained clearly and publicly at the beginning of the experiment.

2.4.1.2 Market experiment

In the labor market, participants play the role of workers competing to be hired by automated firms. Before interacting in the market, instructions about the labor market are distributed to participants and a recording of the instructions is played aloud. Then, the participants answer comprehension questions about the market, including how prices and quantities are determined. Only after the above instructions about market procedures, subjects receive information about the nature of the job. This ensures that subjects in both treatments interpret the market instructions in the same manner. Their understanding of these new instructions about the job is again tested through comprehension questions.

The job. In both markets, workers have the opportunity to be hired as an "advisor" whose job is to give advice to another uninvolved participant outside the laboratory, the "client." Specifically, the advisor has to write a recommendation to a client to choose one among ten choice options (labeled by the letters "A" through "J"). Which option the advisor

[24] The actual cost of providing 30 malaria treatments for children was CHF 29. In order to create small units with a strong moral component, our instructions always referred to the amount of money corresponding to treating "one child" and did not specify the exact monetary amounts.

[25] This implies that Participant As (whose role was not switched) received their own payment with certainty, their decision, however, only had consequences for Participant Bs (and UNICEF) with a probability of 26.3 percent. This corresponds, roughly, to the stochastic impacts in our experimental labor market.

must recommend depends on the treatment. The client receives this recommendation, and then selects one of the ten options. The client only knows that the option he selects determines his financial reward for completing a survey and influences a donation to UNICEF, but does not know the consequences of any specific option. However, the client knows that the advisor had complete payoff information at the time of writing the recommendation. The client is free to choose the recommended option or any other option.[26]

The payoffs associated with each of the ten options are indicated in Table 3. Nine options increase the client's reward by 1 CHF (≈ 1 US Dollar) and increase the donation to UNICEF by an amount estimated to correspond to the anti-malarial treatment of one child. However, one of the 10 options—in this case, option *D*—gives 0 CHF to the client, and *reduces* the donation to UNICEF.

Table 3. Options available to the "client"

	A	B	C	D	E	F	G	H	I	J
Additional number of children who receive the anti-malarial treatment	1	1	1	-1	1	1	1	1	1	1
Financial reward for client	1 CHF	1 CHF	1 CHF	0 CHF	1 CHF	1 CHF	1 CHF	1 CHF	1 CHF	1 CHF

Our two treatments vary only in the recommendation that the advisor is hired to make to the client. In the *neutral treatment*, the advisor's job is to recommend a specific option that is beneficial to the client and to UNICEF (e.g., option *G* in Table 3). Note that a client is very likely to make such a choice independently of any advice, thereby making the impact of such advice largely neutral. In the *immoral treatment*, the job is to recommend the single option with negative consequences (option D). In both cases, the advisor makes a recommendation by completing a form stating that, the option "will save the highest number of children" and "will give you the highest financial reward."[27] By recommending option D, the advisor increases the chance that the client selects the single option that will not increase his earnings and that will reduce the donation to UNICEF. We vary the letter of the recommended bad (neutral) option across immoral (neutral) laboratory sessions. Note that treatments only differ in the moral nature of the job; everything else, including effort costs, is kept constant.

[26] Subsequently, 84% of all recommendations produced in the laboratory were followed by clients.

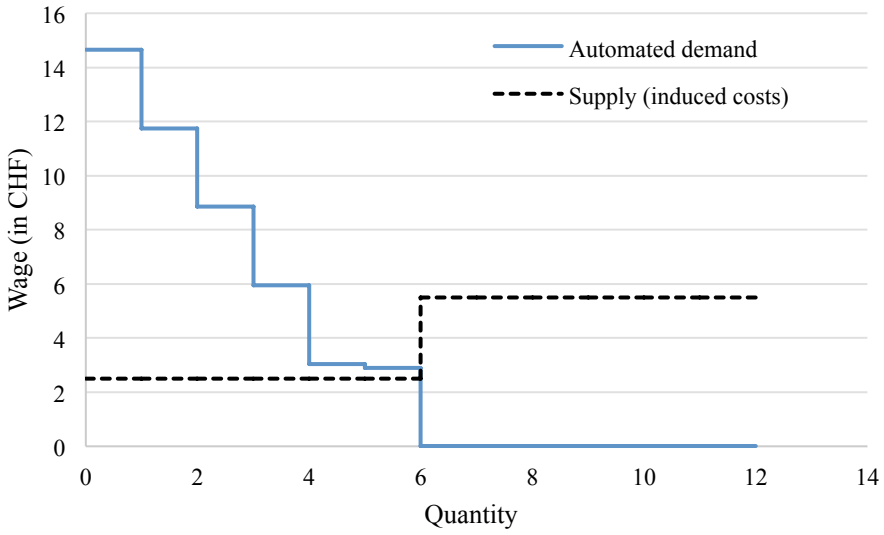
[27] Specifically, the advisor has to complete the following recommendation with the option's name (e.g., G) and his initials: „I, [advisor's initials], have reviewed your possible choices and I recommend that you select the option [G]. Following my advice will save the highest number of children and will give you the highest financial reward. Your advisor: [advisor's initials]"

The market. Participants are randomly allocated to markets consisting of 6 workers who compete to be hired by 6 automated firms. Each worker can provide up to two units of labor—one at a low cost (50 points = CHF 2.50) and one at a high cost (110 points = CHF 5.50). The induced costs are the same for all participants, which the instructions clearly explain.

Each worker decides whether or not to participate in the labor market. In the former case, she (privately) provides two wage requests, one for each of the possible units of labor she can provide. Workers may only submit wage requests that are at least as high as the corresponding cost of providing that job.

Firms are simulated by the computer. Each firm can hire up to one unit of labor per period. Firms are identical except for the wage that they offer to the workers. Figure 2 displays the automated demand for labor as well as the induced costs of labor supply. In equilibrium, all workers provide one unit of labor and the market wage is between CHF 2.5 and 2.9.[28] The workers have no information about the shape of the automated demand.

Figure 2. The automated demand and the induced costs of the labor supply

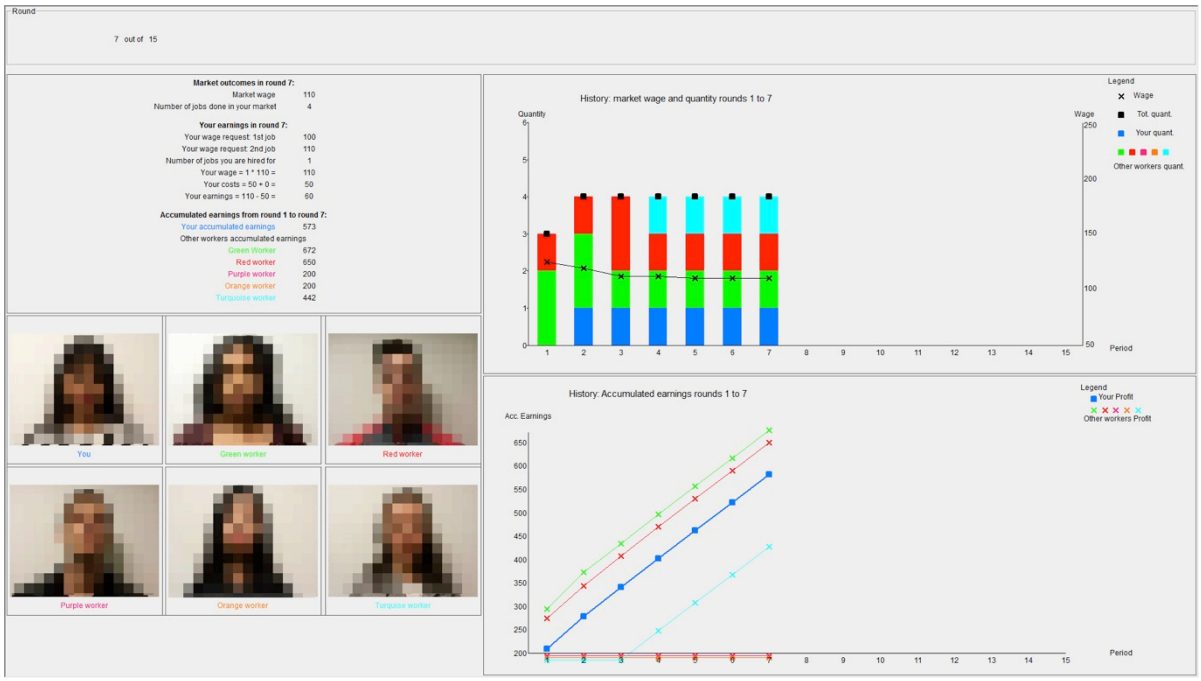


We use a uniform-price sealed-offer auction as the market mechanism, as this provides desirable features. First, Smith et al. (1982) show that this type of market typically converges to the equilibrium prediction. Second, and more importantly, this mechanism allows us to automate the labor demand (see also Sausgruber and Tyran, 2011) and therefore to keep the demand constant between the neutral and the immoral treatments. Once all six

[28] We selected this specific labor demand function to facilitate equilibrium convergence. As long as the wage is substantially higher than the equilibrium prediction (that is, higher than CHF 3.05), at least two workers will be unemployed, putting downward pressure on wages.

workers have submitted their wage requests, the computer ranks them from lowest to highest and compares the workers' wage requests to the firms' wage offers, ranked from highest to lowest. The *market wage* is then the lowest number between two potential candidates: (i) the last wage offer that is higher than the wage request with the same rank, and (ii) the first wage request that is higher than the wage offer with the same rank. This mechanism clears the market in the sense that, for the market wage, labor supply equals labor demand and all workers with wage requests below the market wage are hired.

Figure 3. Example of feedback provided after every period



The market repeats for a total of 15 market periods. The composition as well as the type (immoral or neutral) of each market is fixed across periods. At the end of each period, the computer reports the market wage, displays the picture of every worker in the market and summarizes information regarding each workers' outcomes across all periods (see Figure 3). Specifically, subjects observe employment outcomes, wages and cumulative earnings for all workers in their market across periods, and can connect these to the other workers' identities through the photographs. After observing outcomes, those participants who were hired in a period completed the paper forms with the recommendations—they wrote their own initials and the appropriate letter (e.g., "G" or "D" in the earlier example).[29] If the firm does not

[29] Subjects are informed that in each period each firm has a probability of 25 percent of having a client, which is independent of whether or not the firm hires a worker. If the firm does not have a client, then the worker's recommendation will be unused, although the worker still completes the recommendation and receives the market wage. However, subjects do not know at the time of submitting wage requests or completing the forms

succeed in hiring a worker in a period, the firm's client will not receive any recommendation. This implies that in the immoral treatment, if a participant is not willing to do the job for the market wage, the number of clients who receive bad advices (weakly) decreases.

2.4.1.3 Procedural details

All sessions took place at the Decision Sciences Laboratory (DeSciL) at the Federal Institute of Technology in Zurich (ETH) in February, April and May 2017. Participants were recruited using hroot (Bock, Baetge and Nicklisch, 2014) from the joint subject pool of the University of Zurich and the ETH. Every session consisted of 24 participants, who were only accepted at the session if they had previously completed the on-line survey.[30] To start a session, subjects had to enter an identifier that allows us to link, anonymously, their answers in the on-line questionnaire with their behavior in the lab. We conducted ten sessions, resulting in a total of 240 participants, allocated to 28 immoral markets and 12 neutral markets. The laboratory experiment was implemented with zTree (Fischbacher, 2007).

All instructions were delivered both on paper and with pre-recorded audio files. Instructions and materials are available in the Online-Appendix.[31]

2.4.1.4 Survey study (clients)

We subsequently recruited a different sample of students on the campus of the University of Zurich and the Swiss Federal Institute of Technology (ETH) (N=177). We invited student passersby to participate in a brief choice experiment in which they could earn money and generate a donation for UNICEF aimed at providing treatments for children infected by malaria. They were told that they would earn CHF 2 plus possibly some additional money for a 5-minute study. These subjects performed two functions.

---

at the end of a period whether or not there will be a client for this period. At the end of the experiment, subjects learn which of their written recommendations will be distributed. We did this to lower the number of clients we have to recruit as part of the follow-up survey. This procedure implies that writing a recommendation has only consequences with a probability of 25 percent and, therefore, works against our treatment effect. This represents, for instance, a case in which a worker is hired to prepare promotional materials for a harmful product, which may or may not ultimately be used in a marketing campaign.

[30] We made an exception if less than 24 subjects who completed the survey showed up to the experiment. In total, three subjects were allowed to participate despite not completing the online-survey.

[31] At the conclusion of the laboratory session, we collected several additional individual-level measures. First, we measured participants' affect levels—i.e., pleasure, arousal, and dominance—using the Self-Assessment Manikin (Bradley and Lang, 1994). Next, we asked participants whether they thought the clients would or would not follow recommendations; this belief was not incentivized. The participants were then prompted to answer several questions about the reasons underlying their market behavior. Finally, we measured subjects' concerns for social image using the public self-consciousness scale by Leary et al. (2015), in which participants rate seven short descriptions of behaviors by people who care or do not care about their social-image, on a scale from 1 (*not like me at all*) to 4 (*a lot like me*).

First, they served the role of "clients" for the recommendations from the laboratory labor market. Each participant made up to six decisions by choosing one of the ten letters between A and J. They knew that these decisions influenced their own earnings and also possibly the amount of donations to UNICEF, but they did not know the actual payoffs. Each decision had the payoff structure in Table 3, but we varied which letter corresponded to the bad option. Clients received a mixture of recommendations with good advice, bad advice and no advice (corresponding to the case in which a firm was not able to hire a worker). Clients were only informed of the total payoffs at the end of their decisions.

Second, while their payment was determined and prepared, participants completed a survey in which they rated various firms and industries on a scale from 1 (*very immoral*) to 5 (*very moral*). For firms, clients also had the option to choose "I don't know this organization." The complete list of firms and industries is available in Appendix Tables A1 and A2.

2.4.2 Results

In presenting our results, we focus attention on the existence of an immorality premium and sorting by heterogeneous moral types. We first discuss how we construct our incentivized measure of concern for morality, θ^{Exp}. Next, we study *behavior* in the labor market, and whether it can be predicted by θ^{Exp}. We then study the *outcomes* in the labor market and their connection to the behavioral measure of concern for morality (θ^{Exp}), testing the predictions of our model. While choices in our experiment were incentivized with "points," we present the results in terms of ultimate payments in Swiss francs (CHF) to provide a clearer indication of the economic relevance.

2.4.2.1 Construction of θ^{Exp}

We construct θ^{Exp} based on choices in the behavioral task that subjects completed at the beginning of the laboratory session (Appendix Table A3 shows the distribution of choices). Let m_{ir} be the number that individual i reports if the actual die roll is r. We classify an individual as *low-theta* if $m_{ir} \geq r$ for all $r \in \{1, 2, \ldots, 6\}$ and $m_{ir} > r$ for at least one r; that is, participant i is classified as having a low concern for morality if he or she lies at least once for personal gain and never in self-harmful manner. We classify the remaining participants as *high-theta*.[32] Based on this classification, we have 66 (27.5 percent) low-theta

[32] A total of 13 subjects (5.4 percent, see Appendix Table A3) harmed themselves with a lie ($m_{ir} < r$, e.g., reporting $m_{ir} = 1$ when $r = 2$). Since these subjects do not appear to be motivated by egoism, we classify them as high-theta. The remaining 161 subjects classified as high-theta always report the true number. Classifying

types and 174 (72.5 percent) high-theta types.[33] Given our interpretation of θ, we will often refer to low-theta types as *immoral* and high-theta types as *moral types*. We next explore the differential behavior of the different types in the labor markets and the consequences of this behavior.

2.4.2.2 Labor supply of moral and immoral types

Assuming that θ^{Exp} measures a stable concern for morality that translates into labor-market choices, we should observe differential behavior in the laboratory labor markets between high-theta and low-theta types, but only when employment requires immoral work. This is confirmed in the data. In Table 4, we report the results of a double-hurdle regression of the decision of whether to submit a wage request and, conditionally, the actual wage request. The key independent variable is a subject's type from the behavioral task at the beginning of the experiment. In the immoral treatment, high-theta workers opted to submit wage requests less frequently than low-theta workers (61.6 percent vs. 90.6 percent, $p<0.001$). By declining to submit a wage request, a subject indicates an unwillingness to do the work even at a wage of up to 50 CHF (1000 points), the highest possible wage request in our experiment. Furthermore, consistent with the model, low-theta types submit conditional reservation wage requests that are approximately 0.49 CHF lower than the wage request of high-theta types ($p=0.073$). These effects do not become weaker over time—if anything, the coefficients indicate that the greater willingness of low-theta types to participate in the immoral labor market becomes slightly stronger over time.[34] Moreover, 21.5 percent of the high-theta workers never participated in the market (i.e., refused to submit a wage request in any of the 15 periods), but this is true of only 4.3 percent of low-theta workers ($t = -3.78$; $p=0.001$). Hence, our behavioral measure of a subject's moral type (θ^{Exp}) seems to predict their willingness to seek employment in an immoral job.

---

subjects that lied in a self-harmful manner as high-theta types is conservative in that they act less morally than the honest subjects (see Table A5 in the Appendix). Results do not change if we drop these subjects or if we classify them as low-theta types instead.

[33] In principle, we could classify subjects into more than two categories—e.g., conditional on the number of lies or based on the expected payoff from lying ($\frac{1}{6}\sum_{r=1}^{6} m_{ir} - r$). Due to the low number of subjects with different lying-patterns (see Appendix Table A3), we opt for a binary classification. However, as Appendix Table A5 indicates, we find similar results if we use these alternative classifications.

[34] Specifically, if we add a linear time trend to the hurdle model and its interaction with theta (see Appendix Table A4), we find that low-theta types become slightly more likely to participate over time and provide lower reservation wages, relative to high-theta types. However, both coefficients are small and statistically insignificant.

Table 4. Relationship participation decision/reservation wage and θ^{Exp}

Dependent variable:	Participate	Reservation wage	Reservation wage
	(1)	(2)	(3)
Low-theta (θ_L^{Exp})	1.024***	-0.494*	-0.018
	(4.64)	(-1.79)	(-0.22)
Constant	0.295**	4.056***	2.909***
	(2.41)	(20.20)	(43.79)
Sigma		2.64***	0.609***
		(7.72)	(10.57)
Market	Immoral	Immoral	Neutral
N	2520	1755	1077
LL (pseudo)	-1427.9	-4194.1	-993.9

*Notes: Estimates from Craggs double-hurdle Model: (1) is a probit model; (2) and (3) are truncated linear regressions (truncated from above at 50 CHF). Models (1) and (2) use only data from the immoral markets; model (3) uses only data from the neutral markets. For neutral markets, we do not report the regression of market participation as we have only 3 incidences in which a subject did not participate. Independent variables: Low-theta in {0, 1}. Standard errors clustered at market level; z-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

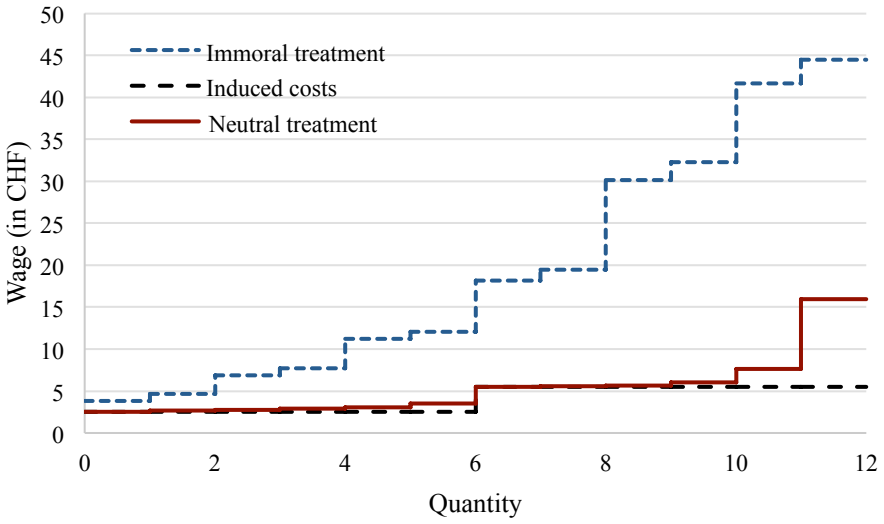
In the neutral treatment, however, non-participation is virtually non-existent—there were only 3 cases in total in which participants chose not to participate, representing 0.28 percent of all observations. That is, there is nearly universal participation when the job does not involve immoral behavior. Moreover, the average wage requests of high-theta (CHF 2.91) and low-theta (CHF 2.89) types do not differ in magnitude or in statistical significance (p-value from hurdle model = 0.827).

Thus, there seems to be differential participation in the market between moral and immoral types, but only when working involves immoral acts. In particular, high-theta types withdraw their participation and make higher wage requests when work requires immoral behavior. However, when the work activity is neutral, both types almost always participate and make similar wage requests. As a direct consequence of these observations, labor supply differs substantially between the two kinds of markets, as shown in Figure 4.[35] In the neutral treatment, labor supply is fairly close to the induced costs. However, for any given wage, there is a substantially lower supply of labor in the immoral treatment.

In the following sections, we explore the implications of the above heterogeneous behavior for labor market outcomes.

[35] Figure A1 in the Appendix shows the labor supply if we only consider the last 5 periods. Figure A2 in the Appendix displays the labor supply in (simulated) labor markets with only low-theta or only high-theta types.

Figure 4. Empirical labor supply for neutral and immoral work in the laboratory



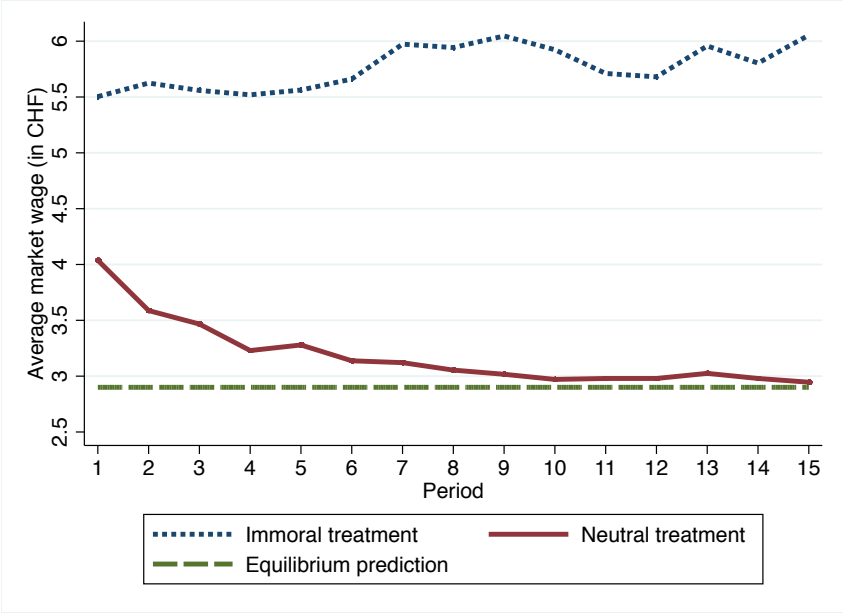
Notes: Wage requests are ranked within each market period of each group. The figure shows the average wage request for each rank for both the immoral and the neutral treatment. Note that wage requests are censored at the maximal wage request that subjects could make, 50 CHF. For this figure, we set the wage requests of subjects who are not willing to participate to CHF 50. Therefore, the supply curve for the immoral treatment should be interpreted as a lower bound.

2.4.2.3 Wage premium in immoral labor markets

In line with Proposition 1, we find a substantial immorality premium, as shown in Figure 5. This follows from the differential labor supply in Figure 4. While market wages in the neutral treatment converge toward the equilibrium prediction of CHF 2.90, the average market wage is persistently higher in the immoral treatment and this difference is statistically significant in a t-test from a regression with standard errors clustered at the market-level (coefficient=2.581, t=6.00, $p<0.001$). Hence, our laboratory labor market yields a substantial and persistent wage premium for immoral work in a setting in which only the morality of work varies. This laboratory evidence corroborates the field evidence in support of Proposition 1 shown in Figure 1 and Table 2.

Of course, one concern might be that the immorality premium is the result of the specific labor demand structure that we employ in our design. However, as evident from the differences in labor supply that we identify in Figure 4, workers are sufficiently concerned with acting morally that we would very likely obtain wage premiums under a wide variety of demand specifications.

Figure 5. Immorality wage premium in laboratory labor markets



We also find persistent differences in the employment levels in the two markets (Figure A3 in the Appendix). While the neutral market converges to the equilibrium prediction of 6, the average market quantity remains below 4 in the immoral markets. This difference is significant in a t-test comparing the means (coefficient=-1.201, t=-6.30, p<0.001). Moreover, the trends in Figure 5 provide further evidence that the manifestation of high-theta participants' morality in labor market behavior does not erode over the course of the experiment. This persistence is remarkable given that participants receive much social information at the end of each market round. Remember that participants learn about past employment and accumulated earnings of all workers in their market. In immoral labor markets, moral participants therefore see other participants less concerned with morality earning high wages due to their own reluctance to act immoral.

2.4.2.4 Sorting in immoral labor markets

We next turn to Proposition 2, which predicts that low-theta types will be disproportionately hired in the immoral markets. Figure 6 shows that, indeed, high-theta types are consistently employed less frequently in the immoral treatment. Table 5 shows that, on average, low-theta types are 26.6 ($= 26.8 - 0.2$) percentage points more likely to be employed than high-theta types (column 1). This difference is highly significant (p<0.001) and robust to adding market fixed effects (column 2). Moreover, this finding is robust to other ways of constructing Θ^{Exp} from behavior in the behavioral task (see Table A5 in the Appendix). The results are similar if we use, as a dependent variable, the number of work units provided (0, 1 or 2) rather than a

binary measure of employment (column 3 and 4). In the neutral treatment, we do not find a significant difference in employment rates between the two types (columns 1 to 4; see also Appendix Figure A4). This corroborates that the difference in hiring rates in the immoral treatment is driven by differences in concerns for morality and not some other difference between high- and low-theta types.

Figure 6. Employment rate by the two types in the immoral treatment

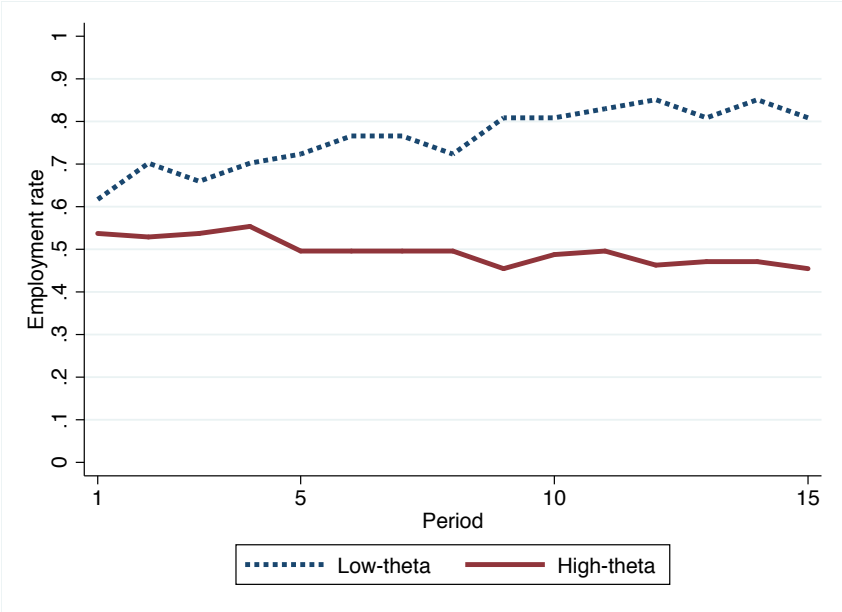


Table 5. Relationship between θ^{Exp} and outcomes in the experimental labor markets

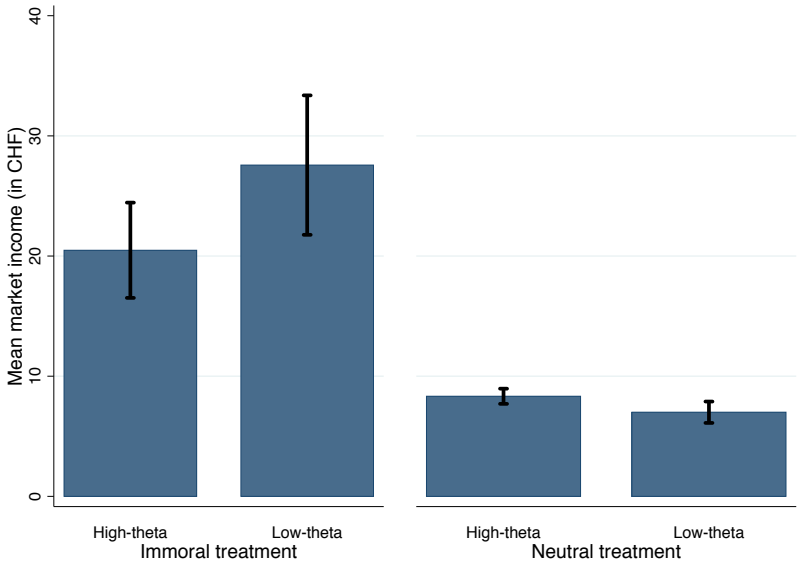
Dependent variable:	**Employment rate**		**Number of jobs**		**Market income**	
	(1)	(2)	(3)	(4)	(5)	(6)
Low-theta (θ_L^{Exp})	-0.002 (-0.05)	-0.034 (-0.90)	-0.002 (-0.05)	-0.034 (-0.90)	-1.33** (-2.40)	-0.56 (-1.39)
Immoral market	-0.331*** (-7.70)		-0.272*** (-7.11)		12.15*** (9.96)	
θ_L^{Exp} * Immoral market	0.268*** (4.54)	0.249*** (3.58)	0.256*** (4.10)	0.264*** (3.57)	8.42*** (2.71)	1.37*** (3.64)
N	240	240	240	240	240	240
R^2	0.179	0.319	0.103	0.174	0.138	0.243
p-value: θ_L^{Exp} + θ_L^{Exp}*Im = 0	0.0000	0.0007	0.0000	0.0009	0.026	0.001
Market FE	No	Yes	No	Yes	No	Yes

*Notes: Coefficient estimates of linear regression models. Independent variables: Low-theta in {0, 1}, Immoral market in {0, 1}. Standard errors clustered at market level; t-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

2.4.2.5 Market income in immoral labor markets

Propositions 3 and 4 predict heterogeneous treatment effects in terms of worker's utility. As a simple proxy for utility, we use the sum of all earnings accumulated by a worker over the 15 market periods. However, as we show in the Appendix C, the same results obtain if we use a slightly more complicated measure that incorporates estimates of workers' disutility from work, which we obtain from their wage requests.

Figure 7. Market income by moral type

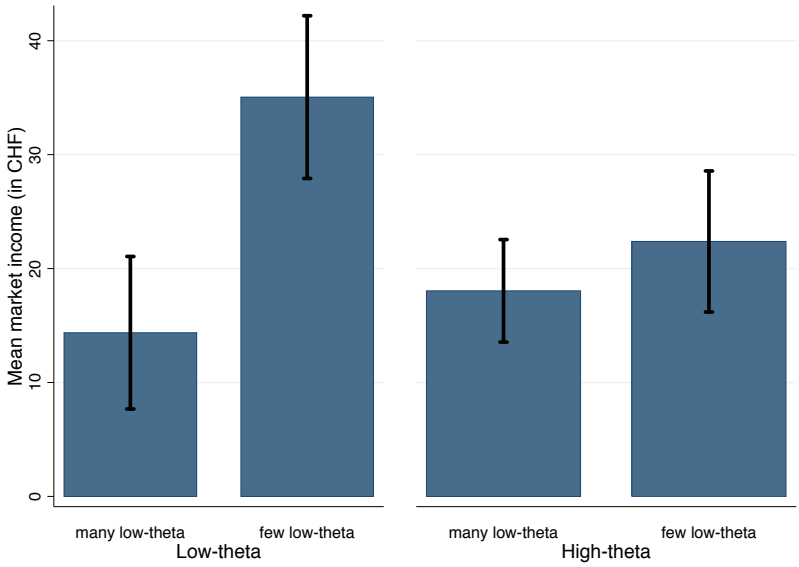


Proposition 3 predicts that the immoral types benefit from an increase in the job's immorality. Figure 7 and results from linear regressions (see Table 5, Columns 5 and 6) show that the low-theta types earn more market income than the high-theta types in the immoral treatment, but not in the neutral treatment. The difference of CHF 7.09 (8.42 – 1.33 in column 5) is statistically significant (p=0.026). Note that the potential market income is constrained by the market wage. If we control for the market wage by adding market fixed effects (column 6), the immoral types are estimated to earn CHF 10.81 more than the moral types (p=0.001). Thus, as predicted by Proposition 3, immoral types earn considerably more in an immoral market. We do not find such a difference in the neutral market; if anything, low-theta types earn slightly less than the high-theta types.

Finally, Proposition 4 predicts that, in the immoral treatment, immoral types have higher utility in the presence of more moral types. To test this prediction, for each subject we count the number of other workers of low-theta type in her market. We then split the sample

based on the median of this measure: we classify a subject as being in a market with *few low-theta types* if the number of (other) low-theta type workers is lower than 2, and as being in a market with *many low-theta types* if the number of (other) low-theta type workers is 2 or more. This results in 70 subjects in the first category, and 98 subjects in the second category. The mean earnings of subjects in the immoral treatment, based on their own type and the median type of others in their market is shown in Figure 8, which uses only data from the immoral treatment. The income of high-theta types is CHF 4.33 higher in a market with few low-theta types than in one with many low-theta types (t=2.05, p=0.051, see Table A6 in the Appendix). For low-theta types, being in a market with few low-theta types increases the income by an additional CHF 16.35 (t=3.53, p=0.002), resulting in a total difference of CHF 20.68 (t=4.29, p<0.001).[36]

Figure 8. Externalities of moral behavior for immoral types



Our laboratory findings confirm all four predictions from our simple model of heterogeneity in the morality of work and concerns for morality. We find evidence that people with high concern for morality consistently refuse to do immoral jobs (or require a high wage), thereby decreasing labor supply and increasing the equilibrium wage, resulting in an immorality premium. As a consequence, subjects with a low concern for morality are

[36] If we use the number of (other) low-theta types in the market instead of doing a median split, we find similar results. For high-theta types, the income is estimated to increase by CHF 2.45 per additional high-theta type in the market (t=2.10, p=0.045). For low-theta types, the income increases by CHF 8.21 for every additional high-theta type (t=2.62, p=0.014).

better off in immoral markets, in particular if they are in a market with many moral subjects.[37]

2.5 Stated real-world employment preferences and sorting

Several days (4-7) before subjects participated in the lab experiment, they filled out an online-survey.[38] The online-survey includes several questions designed to measure subjects' expectations of their own future labor market outcomes, including the willingness to work for different firms and industries and expected future wages, and subjects' concern for morality. The online-survey gives us a second measure of concerns for morality, and, more importantly, allows us to investigate whether subjects' moral types not only predict behavior in the laboratory labor markets but also expectations for real labor market outcomes—in line with Proposition 2.

We first construct an individual measure of concern for morality based on the answers to the psychological survey questions (θ^{Sur}). We then show that this second measure of concern for morality correlates both with the comparable behavioral measure from the laboratory experiment (θ^{Exp}) and with outcomes in the laboratory labor market. This validation of θ^{Sur} is useful for future research, as it is based solely on survey questions which are easier to collect than the incentivized measures. Moreover, the comparison of Θ^{Exp} and θ^{Sur} provides some evidence on the stability of moral concerns across time and contexts, which is necessary for heterogeneous moral concerns to persistently influence labor market behavior.

We then show that both θ^{Sur} and θ^{Exp} predict stated labor market preferences in real labor market, consistent with the sorting process described in Proposition 2.

2.5.1 The on-line questionnaire

We asked subjects several questions about their future labor-market expectations. Subjects were shown a list of 26 well-known companies in Switzerland and another list

[37] One additional consequence of heterogeneous concerns for morality is that the income distribution differs substantially between the two treatments. While in the neutral treatment, income is almost equally distributed (Gini coefficient=0.15), we find substantial income inequality in the immoral treatment (Gini coefficient=0.38). Appendix Figure A5 shows the Lorenz curves for both treatments.

[38] Subjects could only sign up for both the on-line survey and the lab study. Three (out of 240) subjects did not complete the online-survey. We exclude these subjects from this part of the analysis.

consisting of the 20 industries in Figure 1.[39] Both lists are available in Tables A1 and A2 in the Appendix. Participants rated their willingness to work for each firm and industry (1: *not at all willing;* 5: *very much willing*). For firms, participants also had the option to choose, "I don't know this organization," instead of rating their willingness to work for that firm. In addition, we asked subjects to provide unstructured responses stating beliefs about their future career trajectories—specifically, what work they expected to do after their studies and how much they expected to earn at the age of 40.

In addition, we included several multi-item scales intended to measure an individual's broad concern for morality and moral acts. These were:

1) **HEXACO-PI.** We administered 10 items from the short version of the HEXACO Personal Inventory (Ashton and Lee, 2009) related to the factor "Honesty-Humility"—consisting of the four traits, sincerity (3 items), fairness (3 items), greed avoidance (2 items) and modesty (2 items). Every item describes a thought that a moral or immoral person might have and participants indicate the extent to which each thought reflects their own opinions.
2) **Protected Values.** The Protected Values scale (Gibson et al., 2013) measures an individual's position regarding values that can be seen as inviolable, and not substitutable against money, and that are usually central to the person's identity. In our case, and following Gibson et al. (2013), we adapted the Protected Values to a situation where a financial adviser can give bad investment advice to a client for personal benefit. First, 5 items assess the morality of this behavior (*Protected value 1*); second, 4 items examine how truthfulness matters in such a situation (*Protected value 2*).
3) **Integrity and Work Ethics Test.** We used two items from an on-line test designed to allow firms to measure the integrity of job applicants (*Work ethics 1, Work ethics 2*). In each item, participants read fictitious dialogues between two characters with different opinions about a situation (e.g., calling in sick at work to enjoy a sunny day outside). Participants then rate with which character they agree more.
4) **Charity attitude index.** We used a 9-item scale developed by Brashear et al. (2000) in which participants rate statements regarding how important they perceive it is to help others in society and how positive and useful they perceive work done by charities.

[39] Subjects fill out the survey before they learn anything about the rest of the questionnaire or the experiment. Also, note that

In each case, subjects expressed agreement or disagreement with statements on either a 5-point or 7-point Likert scale. Thorough descriptions of these survey scales are provided in the Online Appendix and in Table A7 in the Appendix.

Finally, we asked subjects whether several non-profit organizations (including UNICEF) are worth supporting and collected additional personal characteristics using a short version of the Big Five questionnaire (Gosling et al., 2003), which identifies individuals' extraversion, agreeableness, neuroticism, openness, and conscientiousness, but is largely orthogonal to morality. We also elicited several demographic characteristics, such as age, gender and field of study. The on-line questionnaire was implemented with the Qualtrics software.

2.5.2 Constructing θ^{Sur}

In the following, we discuss how we construct a survey-based measure for concerns for morality, θ^{Sur}. Table 6 lists the 9 subscales from the morality measures we collected. Summary statistics for each of these are provided in Appendix Table A7. We aggregate these nine psychological measures by performing a principal-component factor analysis. We selected the factor with the highest eigenvalue (eigenvalue = 2.44) to represent our *psychological* measure of concern for morality, i.e., θ^{Sur}.[40] Table 6, column 1 presents the corresponding factor loadings. We normalized θ^{Sur} such that it lies between 0 and 1; the resulting variable has a mean of 0.5 and a median of 0.5. Low values represent a low concern for morality. The distribution of θ^{Sur} is presented in Appendix Figure A6. Given our interpretation of θ, we will often refer to subjects with a low Θ^{Sur} as *immoral types* and subjects with high θ^{Sur} as *moral types*.

[40] In the Appendix (Table A8, A9 and A10), we demonstrate that our results are robust to different aggregation mechanisms. Specifically, we look at two alternative aggregation mechanisms: i) each of the nine survey measures is given equal weight and ii) weight of the measures is determined by a regression of θ^{Sur} on the survey measures.

Table 6. Items comprising θ^{Sur} and their relationship to θ^{Exp}

	Factor loadings (weights for θ^{Sur}) **(1)**	**Regression coefficient of θ^{Exp}** **(2)**
Protected value 1	0.664	0.540*** (3.66)
Protected value 2	0.708	0.352** (2.22)
Work ethics 1	0.213	-0.039 (-0.39)
Work ethics 2	0.252	0.042 (0.52)
HEXACO sincerity	0.482	0.362** (2.53)
HEXACO fairness	0.611	0.353*** (2.61)
HEXACO greed avoidance	0.477	0.225* (1.73)
HEXACO modesty	0.508	0.236* (1.79)
Charity attitude index	0.545	0.711*** (3.11)

*Notes: Each subscale is constructed by taking averages over all items of the scale, and then normalized such that it lies between 0 and 1. (1): Factor loadings from principal-component factor analysis of survey measures on θ^{Sur}. (2): Coefficient estimates of linear probability models. N = 237 for each regression (3 subjects did not complete the online-survey and are excluded). Dependent variable: being a high-theta type according to θ^{Exp}. Independent variables: survey measures in [0,1], higher numbers indicate more morality. Robust standard errors; t-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

2.5.3 Does θ^{Sur} predict behavior in the laboratory?

To validate θ^{Sur}, we investigate how it correlates with behavior in the laboratory. Table 6, column 2, shows the coefficients from independent simple regressions of a subject's type measured by the behavioral laboratory task, θ^{Exp}, on all items comprising θ^{Sur}. The dependent variable is binary, indicating that a subject is a high-theta type. The results show a significant positive correlation between θ^{Exp} and all personality measures, except for *Work ethics 1* and *Work ethics 2*. Consistent with the positive relationship of the individual items, a regression of θ^{Exp} on θ^{Sur} shows positive and significant relationships (coefficient=0.723, t=4.32, p<0.001), that is, a person who is characterized by a low concern for morality according to our survey-based measures is more likely to lie self-servingly in the behavioral measure in the experiment. These generally positive relationships suggest that θ^{Exp} and our measures of psychological traits capture similar individual characteristics.

We next consider the extent to which θ^{Sur} also predicts participants' behavior in the labor market experiment, and particularly in the immoral condition.[41] Results from a linear regression of the employment rate on θ^{Sur} indicate that the participants with the lowest concerns for morality (that is, participants with $\theta^{Sur} = 0$) are 43.9 percentage points more likely to be hired in immoral labor markets than the participants with the highest concerns for morality (that is, participants with $\theta^{Sur} = 1$) and this difference is marginally statistically significant (p=0.057, see Appendix Table A8, column 1). A less noisy measure of subjects' market behavior is their actual choices. Results from a hurdle model indicate that the subjects with the lowest concerns for morality are 52.1 percentage points more likely to participate in immoral labor markets by submitting a wage request than the individuals classified as having the highest concerns for morality (p=0.015, see Appendix Table A9). However, they do not have significantly lower reservation wages than the moral types in those markets (p=0.548). In the neutral treatment, as expected, we do not find a significant difference in employment rates or in labor market behavior between the two types.

2.5.4 Do θ^{Sur} and θ^{Exp} predict stated real-world labor market preferences?

Our study collects two measures of participants' concern for morality (θ^{Sur} and θ^{Exp}). Furthermore, the on-line survey elicits the same participants' willingness to work for several firms and industries, without making any reference to morality. We also separately obtained independent ratings of the perceived immorality of these firms and industries, from the "clients." In this section, we use all of this information to analyze how our measures of concern for morality connect to expectations about labor market outcomes outside the laboratory.

We create a measure of *perceived firm immorality* in the same way as we created *perceived industry immorality:* by averaging the ratings and scaling them such that they lie between -1 (very moral) and +1 (very immoral), where 0 means neutral.[42] We use these variables as noisy measures of the immorality of work, $I(j)$, in industry (or, firm) j, a key component of our theoretical model. The horizontal axes in Figures 9a and 9c plot the resulting normalized ratings for industries in our sample, the horizontal axes in Figures 9b and 9d plot the normalized ratings for firms (see also Appendix Tables A1 and A2).

[41] Figure A7 in the Appendix displays the labor supply in a (simulated) labor market with only low-θ^{Sur} or only high-θ^{Sur}types.

[42] Remember that for *firms*, clients also had the option to choose "I don't know this organization" instead of rating the firm. To calculate the perceived firm immorality, we exclude these observations. Alternatively, we could code these as neutral ratings. These two measures are highly correlated (corr=0.9854). Our results do not change substantially if we use the alternative measure (see Table A11).

Our focus in this section is to investigate whether these perceptions of immorality of industries and firms interact with our subjects' measured concern for morality (either θ^{Sur} or θ^{Exp}) to produce differential labor market preferences. For this purpose, we normalized participants' stated willingness to work for firms and industries, such that they take values between 0 (*not at all willing*) and 1 (*very much willing*).[43]

The vertical axes of Figures 9a and 9c plot the difference in willingness to work for the industries between participants who were classified as moral or immoral, according to θ^{Exp} (Figure 9a) or θ^{Sur} (Figure 9c). The strong negative relationship indicates that participants classified as immoral are, on average, more willing to work for industries perceived as immoral.

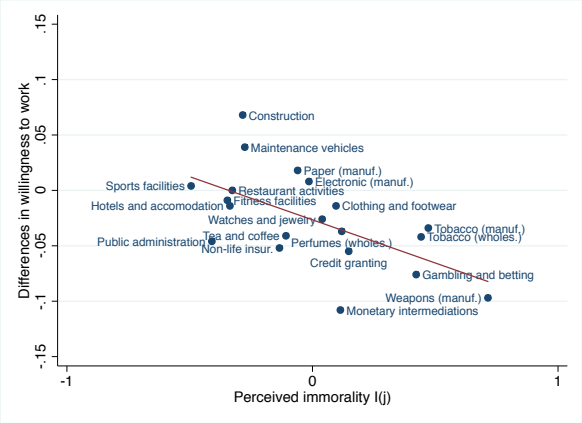
Table 7, columns (1) to (4) provide statistical evidence for the relationships in Figure 9. While there is little evidence for a systematic difference in willingness to work for neutral industries between moral and immoral types, participants classified as immoral are significantly more willing to work in industries that are classified as immoral. This pattern is significant at least at the 5%-level, holds for both measures of individual moral concerns, θ^{Sur} and θ^{Exp}, and is robust to controlling for participants' gender, age, Swiss nationality, area of study, mean industry wages, industry size (number of employees), and industry sales.

We repeat the same analysis using data on individuals' willingness to work for our selection of well-known *firms* in Switzerland. The vertical axis of Figures 9b and 9d plot the difference in willingness to work for the firms between participants who were classified as moral and immoral according to θ^{Exp} (Figure 9b) or Θ^{Sur} (Figure 9d). Again, participants that are classified as immoral are, on average, more willing to work for firms perceived as immoral. This relationship is confirmed by Table 7, columns (5) to (8): immoral participants are more willing to work for firms that other people rate as more immoral ($p<0.01$), which is again true for both measures of concern for morality (θ^{Sur} and θ^{Exp}). This finding indicates that firms that are perceived as immoral differentially attract applicants with a lower concern for morality.[44]

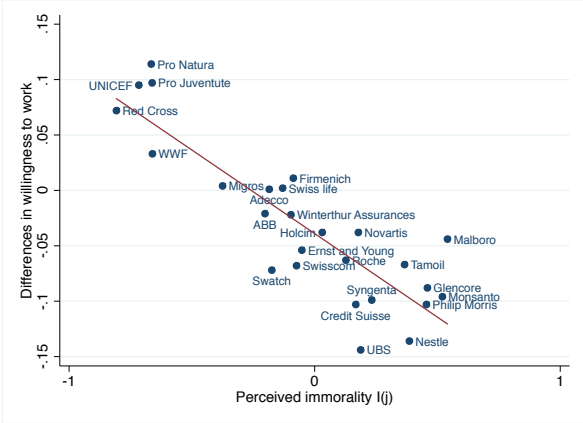
[43] Incidentally, the list of industries accidentally omitted five industries for five participants that participated in the first lab session. So, we are missing data on willingness to work in these industries for these participants. Other than these cases, all subjects completed the full questionnaire. We exclude all these missing observations from the analysis. Regarding willingness to work for *firms*, participants also had the option to choose "I don't know this organization." This option was chosen in 17.8 percent of all answers. We also exclude these observations. We obtain similar results if we classify such observations as "indifferent" or if we restrict our analysis to subjects that know all firms (see Table A12).

[44] As we show in the Appendix, the differential willingness to work for immoral firms and industries by moral and immoral types does not depend on how we construct Θ^{Sur} (Table A10) and how we deal with missing observations (Table A11, Table A12).

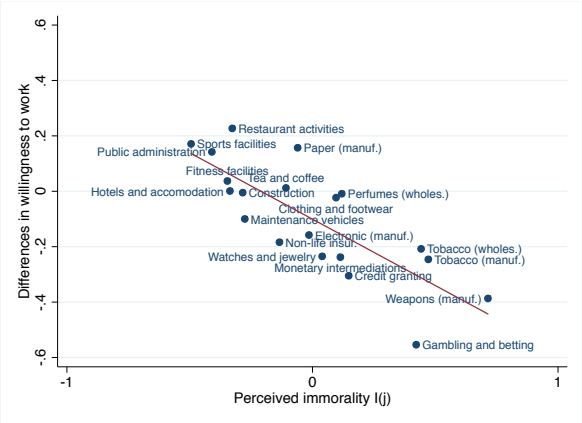
Figure 9. Correlation between the difference in willingness to work between moral and immoral types and perceived immorality of industries/firms.



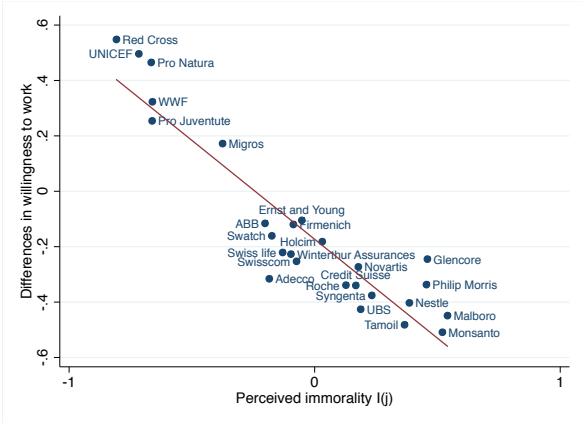
(a) Industries, θ^{Exp}



(b) Firms, θ^{Exp}



(c) Industries, θ^{Sur}



(d) Firms, θ^{Sur}

Source: Survey study (Perceived immorality), on-line survey (Willingness to work, θ^{Sur}), Laboratory experiment (θ^{Exp}).

Notes: Differences in willingness to work: Coefficient estimates of linear regression models of the participants' willingness to work for different industries (a and c) or firms (b and d) on θ_H^{Exp} (a and b) or θ^{Sur} (c and d). Dependent variable: Willingness to work is in {0, 0.25, 0.5, 0.75, 1} where 0 means not at all willing to work, 0.5 means indifferent and 1 means really much willing to work. Observations where subjects did not know the firm ("I don't know this organization") or did not fill out the questionnaire are excluded. Independent variables: a and b use θ^{Exp} to classify participants, where θ_H^{Exp}=0 for low- theta types and θ_H^{Exp}=1 for high-theta types, while c and d use θ^{Sur} in [0,1] instead. Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral.

Table 7: Regressions of willingness to work for diverse industries and firms on perceived immorality and moral types.

Dependent variable:	**Willingness to work for industry *j***				**Willingness to work for firm *j***			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Perceived immorality (I(j))	-0.232*** (-3.99)	-0.226*** (-4.72)	-0.050 (-0.77)	-0.043 (-0.74)	-0.140** (-2.08)	-0.139** (-2.07)	0.114 (1.56)	0.110 (1.50)
Type experiment (θ_H^{Exp})	-0.027 (-1.34)	-0.029 (-1.49)			-0.043 (-1.58)	-0.051** (-1.96)		
θ_H^{Exp} * I(j)	-0.078*** (-2.56)	-0.078** (-2.51)			-0.154*** (-4.30)	-0.154*** (-4.38)		
Type survey (θ^{Sur})			-0.101* (-1.71)	-0.107* (-1.72)			-0.173** (-2.28)	-0.211*** (-2.85)
θ^{Sur} * I(j)			-0.479*** (-5.22)	-0.479*** (-5.24)			-0.731*** (-8.80)	-0.722*** (-8.53)
N	4715	4715	4715	4715	5064	5064	5064	5064
Control variables	No	Yes	No	Yes	No	Yes	No	Yes

*Notes: Coefficient estimates of linear regression models. Dependent variable: Willingness to work is in {0, 0.25, 0.5, 0.75, 1} where 0 means not at all willing to work, 0.5 means indifferent and 1 means really much willing to work. Observations where subjects did not know the firm ("I don't know this organization") or did not fill out the questionnaire are excluded. Independent variables: (1), (2), (5) and (6) use θ_H^{Exp} to classify participants, where θ_H^{Exp}=0 for low- theta types and θ_H^{Exp}=1 for high-theta types, while (3), (4), (7) and (8) use θ^{Sur} (in [0,1]) instead. Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral. Control variables: age, gender, Swiss nationality, subject of study, average wage industry 2016 (SLFS; only for industries), industry size 2016 (STATENT; only for industries), industry sales 2015 (Value Added Tax Statistics; only for industries). Standard errors clustered at individual and industry/firm level (Cameron, Gelbach and Miller, 2011); z-statistics in parentheses; * p < 0.1; ** p < 0.05; *** p < 0.01.*

The above analysis provides evidence supporting Proposition 2 in real labor markets. Those who are least concerned with morality are significantly more willing to work in firms that are perceived as less moral. A limitation of this survey evidence is that it is based on hypothetical future choices. However, Wiswall and Zafar (2018) provide evidence that such stated preferences are predictive of ultimate employment. To further validate subjects' stated real-world labor market preferences, we can test whether the stated employment preferences correlate with individual employment rates in the immoral treatment of our laboratory experiment. Indeed, we find that people that are hired more often for the immoral job have a statistically significant higher willingness to work in immoral industries (see Table A13 in the Appendix).

Finally, in the on-line survey, we also asked participants to rate how much they expect to earn when they reach the age of 40. We do not find any statistically significant correlation between participants' type and their earnings' expectations, although a regression of expected future wages on θ^{Sur} reveals a positive relationship such that the least moral types ($\theta^{Sur} = 0$) report expected income that is 30,272 CHF higher on average than the expected income of the most moral type ($\Theta^{Sur} = 1$, p=0.125). This relationship is consistent with Proposition 1. However, the earnings expectations measures for such a long time

horizon—on average 18 years—are perhaps less reliable than the more contemporaneous statements of willingness to work for different firms.

2.6 Discussion and Conclusion

We investigate how individual heterogeneity in concerns for morality interact with heterogeneity in the perceived immorality of work to influence the types of jobs that individuals select and their earnings. Our study employs a combination of a laboratory experiment, surveys and empirical data to identify heterogeneity in concerns for morality and to measure (or, create) variation in the immorality of jobs. We use these different kinds of data to test two main hypotheses—first, that jobs that are generally perceived as immoral will yield an immorality wage premium and, second, that individuals less concerned with moral behavior will be more likely to be hired in such jobs.

In a laboratory setting, we use a simple behavioral task to classify individuals into "moral" and "immoral" types. We then show that this characteristic predicts the outcomes that individuals obtain as we experimentally vary *only* the immorality of work. We find support for both our hypotheses. Immoral labor markets yield significantly higher wages. Moreover, immoral workers are significantly more likely to be hired in a labor market for immoral work than are moral workers; however, this difference disappears in a labor market for neutral work. We also find that a market for immoral work benefits the immoral types who are hired, particularly when there are many moral types in their market.

We separately use survey responses to classify the immorality of real-world firms and industries and show that industries classified as immoral pay higher wages. We also use surveys to obtain a separate measure of workers' moral types. This individual characteristic is correlated with the moral type measured in the laboratory and predicts subjects' behaviors in the laboratory labor market. Moreover, both the survey-based measure and lab-based measure of morality also predict stated preferences for working in jobs and industries that vary in their morality. Workers who are less concerned with morality—in either the behavioral or survey-based measures—are more willing to work for firms that others regard as less moral.

Despite widespread intuition, we know of no other evidence that makes the above connections. Given the significance of many social ills produced by immoral work practices, such as deceptive marketing or socially harmful products, our study sheds important new

light on the interaction between individual's types, their willingness to do immoral work and the resulting labor-market outcomes.

Our work also has several potentially important policy implications. For instance, in those jobs and industries with the greatest potential to do societal harm, social welfare will often be higher when workers in such industries voluntarily internalize the negative impacts of their actions and forgo potentially profitable opportunities. For instance, a weapons manufacturer may restrict sales to conflict areas if top management has a moral aversion to the social harm caused by such sales. However, our evidence suggests that it is the *least* moral types who will sort into these industries and that, therefore, labor market sorting will make it less likely that such internalization will occur.

Another implication of our empirical findings is that the perception that a firm, industry or type of work is immoral may be self-reinforcing. If, as our results indicate, the perception that work involves immoral acts leads people less concerned with acting immorally to differentially opt into such work, then the end result of such sorting may be a workforce of people who are more likely to commit immoral acts. Even if some of the firms and industries that we study do not actually involve any inherently immoral activities in their line of work, the fact that they disproportionately attract people more willing to do immoral things may mean a greater prevalence for immoral behavior. Thus, firms and industries that regularly confront the perception that they involve immoral work—such as the banking sector—may need to be particularly attuned to such selection in their hiring.

An additional implication of our study is that negative externalities that result from the immoral behavior of firms can be partly internalized by means of an increase in labor costs. That is, increasing the perception that a firm is immoral will lead it to face higher labor costs, thereby creating a partial internalization of the harm it produces. Thus, public discourse and narratives about the ills produced by different firms or industries may be valuable not only for producing change in policy and consumer behavior, but also for influencing outcomes through changes in labor market behavior. However, increasing the labor costs for immoral work in this manner may have the effect of further reducing the average morality of firms' employees. As the work becomes perceived as more morally aversive, only the least moral types will opt into such work and, as we note above, those may also be the types least inclined to internalize any negative externalities.[45]

[45] Relatedly, it has recently been suggested to use taxes to reallocate skilled labor from industries that produce negative externalities to industries important for society (Lockwood, Nathanson and Weyl, 2017; see also Murphey, Shleifer and Vishny, 1991; Rothschild and Scheuer, 2016). While such an intervention would

Finally, our model predicts—in line with our experimental data—that the least moral types are overcompensated by the immorality premium. This is in stark contrast to Mankiw's (2010) "just deserts theory"—that is, everybody should receive his contribution to society. Our work suggests a perverse case in which those willing to do the most socially harmful acts may instead benefit from doing so, particularly as others find this work more aversive.

Of course, our work leaves open many important questions regarding the precise characteristics that lead some kinds of work to be differentially perceived as immoral and the specific nature of the preference underlying workers' market behavior. Nevertheless, as the above examples make clear, the differential sorting by people more or less concerned with immoral behavior into different lines of work has important implications for the extent of which market activity yields beneficial social outcomes.

---

decrease the number of employees in immoral industries, the downside of such a policy is that, according to our theory, the average morality of employees in immoral industries would decrease further.

References

Andreoni, J., Nikiforakis, N. and Stoop, J. (2017) "Are the Rich More Selfish than the Poor, or Do They Just Have More Money? A Natural Field Experiment," NBER Working Paper No. 23229.

Ariely, D., Bracha, A. and Meier, S. (2009) "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," American Economic Review, 99(1): 544-555.

Arunachalam, R. and Shah, M. (2008) "Prostitutes and Brides?," American Economic Review, Papers and Proceedings, 98(2): 516-522.

Ashton, M.C. and Lee, K. (2009) "The HEXACO-60: A Short Measure of the Major Dimensions of Personality," Journal of Personality Assessment, 91(4): 340-345.

Bartling, B., Valero, V., and Weber, R. A. (2018) "Is Social Responsibility a Normal Good?," working paper.

Bartling, B., Weber, R. A. and Yao, L. (2015) "Do Markets Erode Social Responsibility?," Quarterly Journal of Economics, 130(1): 219-66.

Bates, Cl. and Rowell, A. (1998) "Tobacco explained: The truth about the tobacco industry ... in its own words." Scotland: ASH.

Becker, G. S. (1957) "The Economics of Discrimination," Chicago: University of Chicago Press.

Benedict, M.E., McClough, D and McClough, A.C. (2006) "The price of morals: An empirical investigation of industry sectors and perceptions of moral satisfaction – do business economists pay for morally satisfying employment?," The American Economist, 50(1): 21-36.

Besley, T. and Ghatak, M. (2005) "Competition and Incentives with Motivated Agents," American Economic Review, 95(3): 616-636.

Blitz, D. and Fabozzi, F. J. (2017) "Sin Stocks Revisited: Resolving the Sin Stock Anomaly," 44(1): 1-7.

Bock, O., Baetge, I., and Nicklisch, A. (2014) „Hroot: Hamburg registration and organization on-line tool," European Economic Review, 71: 117-120.

Bradley M.M. and Lang, P.J. (1994) "Measuring emotion: the self-assessment manikin and the semantic differential," Journal of Behavioral Therapy and Experimental Psychiatry, 25(1): 49-59.

Brashear, G., Green, L., and Webb, J. (2000) "Development and validation of scales to measure attitudes influencing monetary donations to charitable organizations," Journal of the Academy of Marketing Science, 28(2): 299-309.

Cameron, A. C., Gelbach, J. G., and Miller, D. L. (2008) "Bootstrap-Based Improvements for Inference with Clustered Errors," Review of Economics and Statistics, 90(3): 414-427.

Cameron, A. C., Gelbach, J. B. and Miller, D. L. (2011) "Robust inference with multiway clustering," Journal of Business & Economic Statistics, 29(2): 238-249.

Carpenter, J. and Gong, E. (2016) "Motivating Agents: How Much Does the Mission Matter?," Journal of Labor Economics, 34(1): 211-236.

Carpenter, J., Matthews, P. and Robbett, A. (2017) "Compensating Differentials in Experimental Labor Markets," Journal of Behavioral and Experimental Economics, 69: 50-60.

Carter, J. R., and Irons, M. D. (1991) "Are Economists Different, and If So, Why?" Journal of Economic Perspectives, 5(2): 171-177.

Case, A. and Deaton, A. (2015) "Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century," Proceedings of the National Academy of Sciences, 112(49): 15078-15083.

Cassar, L. and Meier, S. (2018) "Nonmonetary Incentives and the Implications of Work as a Source of Meaning," Journal of Economic Perspectives, 32(3): 215-238.

Cassar, L. (2019) "Job Mission as a Substitute for Monetary Incentives: Benefits and Limits," Management Science, 65(2): 896-912.

Cohn, A., Fehr, E. & Maréchal, A. (2014) "Business culture and dishonesty in the banking industry," Nature, 516: 86-89.

Colonnello, S., Curatola, G. and Gioffré, A. (2019) "Pricing Sin Stocks: Ethical Preference vs. Risk Aversion," European Economic Review, 118: 69-100.

Dal Bó, E., Finan, F. and Rossi, M. A. (2013) "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service," Quarterly Journal of Economics, 128(3): 1169-1218.

Delfgaauw, J. and Dur, R. (2008) "Incentives and Workers' Motivation in the Public Sector," Economic Journal, 118(525): 171-191.

Deseranno, E. (2019) “Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda” American Economic Journal: Applied Economics, 11(1): 277-317.

Dur, R. and van Lent, M. (2019) “Socially Useless Jobs,” Industrial Relations, 58(3), 543-546.

Dur, R. and Zoutenbier, R. (2014) “Working for a Good Cause,” Public Administration Review, 74(2): 144-155.

Edlund L. and Korn, E. (2002) “A Theory of Prostitution,” Journal of Political Economy, 110(1): 181-214.

Edlund, L., Engelberg, J. and Parsons, C. A. (2009) “The Wages of Sin,” Columbia Discussion Paper Series, Discussion Paper 0809-16.

Eriksson, T. and Kristensen, N. (2014) “Wages or Fringes? Some Evidence on Trade-Offs and Sorting,” Journal of Labor Economics, 32(4): 899-928.

Fabozzi, F. J., Ma, K. C. and Oliphant B. J. (2008) “Sin stock returns,” Journal of Portfolio Management, 35(1): 82-94.

Falk A. and Szech, N. (2013) ‘‘Morals and Markets,’’ Science, 340: 707-711.

Fehrler, S. and Kosfeld, M. (2014) “Pro-Social Missions and Worker Motivation: An Experimental Study,” Journal of Economic Behavior & Organization, 100: 99-110.

Fischbacher, U. (2007) “z-Tree: Zurich toolbox for ready-made economic experiments,” Experimental Economics, 10(2), 171-178.

Fisman, R., Jakiela, P., Kariv, S. and Markovits, D. (2015) “The distributional preferences of an elite,” Science, 349(6254): 1300.

Frank, R. H. (1996) “What prices the moral high ground?,” Southern Economic Journal, 63(1): 1-17.

Frank, R. H., Gilovich, T. and Regan, D. T. (1993) “Does Studying Economics Inhibit Cooperation?,” Journal of Economic Perspectives, 7(2): 159-171.

Garen, J. (1988) “Compensating Wage Differentials and the Endogeneity of Job Riskiness,” Review of Economics and Statistics, 70(1): 9-16.

Gertler, P., Shah, M. and Bertozzi, S. M. (2005) “Risky Business: The Market for Unprotected Commercial Sex,” Journal of Political Economy, 113(3): 518-550.

Gibson, R., Tanner, C. and Wagner, A. (2013) “Preferences for truthfulness: Heterogeneity among and within individuals,” American Economic Review, 103(1): 532-548.

Gneezy, U., Rockenbach, B. and Serra-Garcia, M. (2013) “Measuring lying aversion,” Journal of Economic Behavior and Organization, 93: 293-300.

Gosling, S.D., Rentfrow, P.J., and Swann Jr. W.B. (2003) "A very brief measure of the Big-Five personality domains," Journal of Research in Personality, 37: 504-528.

Gregg, P., Grout, P. A., Ratcliffe, A., Smith, S. and Windmeijer, F. (2011) "How important is pro-social behaviour in the delivery of public services?," Journal of Public Economics, 95: 758-766.

Hanna, R. and Wang, S. (2017) "Dishonesty and Selection into Public Service: Evidence from India," American Economic Journal: Economic Policy, 9 (3): 262-290.

Heath, D. (2016) "Contesting The Science of Smoking," The Atlantic, https://www.theatlantic.com/politics/archive/2016/05/low-tar-cigarettes/481116/

Heidhues, P., Kőszegi, B. and Murooka, T. (2016) "Inferior products and profitable deception" Review of Economic Studies, 84(1): 323-356.

Hill, C. A. (2012) "Bankers behaving badly? The limits of regulatory reform," Review of Banking and Financial Law, 31: 675-691.

Hong, H. and Kacperczyk, M. (2009) "The Price of Sin: The Effects of Social Norms on Markets," Journal of Financial Economics, 93: 15-36.

Hwang, H., Reed, W. R., and Hubbard, C. (1992) "Compensating Wage Differentials and Unobserved Productivity," Journal of Political Economy, 100(4): 835-858.

Jones, D. B. (2015) "The Supply and Demand of Motivated Labor: When Should We Expect to See Nonprofit Wage Gaps?," Labour Economics 32: 1-14.

Kirchler, M., Huber, J., Stefan, M. and Sutter, M. (2016) "Market design and moral behavior," Management Science, 62: 2615-2625.

Leary, M.R., Diebels, K. J. and Jongman-Sereno, K. P. (2015) "Measures of concerns with public image and social evaluation", in: Measures of Personality and Social Psychological Constructs, Elsevier Inc., 448-473. DOI: 10.1016/B978-0-12-386915-9.00016-4.

Leete, L. (2001) "Whither the Nonprofit Wage Differential? Estimates from the 1990 Census," Journal of Labor Economics 19(1): 136-170.

Levitt, S. D. and List, J. A. (2007) "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?," Journal of Economic Perspectives, 21(2): 153-174.

Levitt, S. D. and List, J. A. (2008) "Homo economicus Evolves," Science, 319: 909-910.

Lockwood, B. B., Nathanson, C. G. and Weyl, E. G. (2017) "Taxation and the Allocation of Talent," Journal of Political Economy, 125(5): 1635-1682.

Maestas, N., Mullen, K. J, Powell, D., von Wachter, T. and Wenger, J. (2018) "The Value of Working Conditions in the United States and Implications for the Structure of Wages," working paper.

Mankiw, G. N. (2010) "Spreading the Wealth Around: Reflections Inspired by Joe the Plumber," Eastern Economic Journal, 36: 285-298.

Mas, A., and Pallais, A. (2017) "Valuing Alternative Work Arrangements," American Economic Review, 107(12): 3722-3759.

Mocan, N. H. and Tekin, E. (2003) "Nonprofit Sector and Part-Time Work: An Analysis of Employer-Employee Matched Data on Child Care Workers," Review of Economics and Statistics 85(1): 38-50.

Moffatt, P. G. and Peters, S. A. (2004) "Pricing personal services: An empirical study of earnings in the UK prostitution industry," Scottish Journal of Political Economy 51(5): 675-690.

Murphy, K., Shleifer, A. and Vishny, R. (1991) "The allocation of talent: Implications for growth," Quarterly Journal of Economics, 106(2): 503-530.

Okie, S. (2010) "A flood of opioids, a rising tide of deaths," New England Journal of Medicine, 363(21): 1981-1985.

Prendergast, C. (2007) "The Motivation and Bias of Bureaucrats," American Economic Review, 97(1): 180-196.

Pörtner, C. C., Hassairi, N., and Toomim, M. (2015) "Only if You Pay Me More: Field Experiments Support Compensating Wage Differentials Theory," working paper.

Rosen, S. (1986) „The theory of equalizing differences", Ed: Ashenfelter, O. and Layard, R., in: Handbook of Labor Economics, Elsevier Science Publishers BV.

Rothschild, C. and Scheuer, F. (2016) "Optimal Taxation with Rent-Seeking," Review of Economic Studies, 83: 1225-1262.

Ruhm, C. J. and Borkoski, C. (2003) "Compensation in the Nonprofit Sector," Journal of Human Resources 38(4): 992-1021.

Sausgruber, R. and Tyran, J-R. (2011) "Are we taxing ourselves? How deliberation and experience shape voting on taxes", Journal of Public Economics, 95: 164-176.

Sjöberg, L. and Engelberg, E. (2009) "Attitudes to Economic Risk Taking, Sensation Seeking and Values of Business Students Specializing in Finance," Journal of Behavioral Finance, 10(1): 33-43.

Smith, A. (1776) "An Inquiry into the Nature and Causes of the Wealth of Nations," London: W. Strahan and T. Cadell.

Smith, V.L., Williams A.W., Bratton W.K. and Vannoni, M.G. (1982) “Competitive market institutions: Double auctions vs. sealed bid-offer auctions”, American Economic Review, 72(1): 58-77.

Tonin, M. and Vlassopoulos, M. (2015) „Corporate Philanthropy and Productivity: Evidence from an Online Real Effort Experiment,“ Management Science, 61(8): 1795-1811.

US Department of Justice, Office of Public Affairs (2016) “Goldman Sachs Agrees to Pay More than $5 Billion in Connection with Its Sale of Residential Mortgage Backed Securities,” https://www.justice.gov/, date accessed: 21.01.2017.

Wiswall, M. and Zafar, B. (2018) “Preference for the Workplace, Investment in Human Capital, and Gender,” Quarterly Journal of Economics, 133(1): 457-507.

Chapter 3: Consumption, Moral Values and Identity Signaling

Abstract

Major models of identity signaling and consumption postulate that consumers care about the characteristics of others that buy a product. This paper studies whether the type-composition of a product's customer pool indeed influences the demand for the product. In two laboratory experiments, I vary, by treatment, whether the customer pool of a product consists of people with desirable or undesirable characteristics, in particular moral values. I do so in a setting where choices are publicly observable and, as a result, can serve as social-signals of moral types, and in a double blind setting. In both studies, I find that participants are willing to pay more for a product if its customer pool has desirable moral values.

Citation

Schneider, F. H. (2020) "Consumption, Moral Values and Identity Signaling," working paper.

3.1 Introduction

A typical assumption in economics is that firms maximize their profits and that they do so by selling their products to everyone that is willing to pay their prices. However, firms' reactions to product adaptations by alt-rights, neo-Nazis and Hooligans show that many firms want some types of customers *not* to buy their products. For example, after the neo-Nazi website Daily Stormer claimed Papa John's to be the official pizza brand of the alt-right, the company stated that "we do not want these individuals or groups to buy our pizza" (Washington Post, 2017).[46] Firms are willing to bear substantial costs to avoid right-wing extremists as customers. When the clothing brand Lonsdale purposely broke with their neo-Nazi customers, they lost 35 percent of their sales volume in Germany (Handelsblatt, 2014). Why do firms try to avoid some customers? Firms striving for customers not to buy their products seems puzzling from the perspective of profit maximization. Moreover, firms spend substantial amounts of money to attract customers with desirable characteristics by celebrity endorsement and by hiring influencers. This paper gives a potential rationale for these firm behaviors by providing evidence that consumers care about the characteristics of others that buy a particular product. The evidence is consistent with models of identity signaling: consumers signal their desirable characteristics (or, "types") to themselves and others by avoiding products popular among people with undesirable characteristics and by conforming to product choices of people with desirable characteristics. One important implication of this identity signaling motive is that firms have incentives to avoid (or, attract) customers, depending on the social desirability of their characteristics.

Understanding such (non-)conformity is important beyond the context of firms managing the composition of their customer pool. A longstanding theoretical literature in economics, philosophy and sociology proposes that customers signal desirable characteristics through consumption choices (starting with Veblen, 1899/1994; Simmel, 1904/1957;

[46] Wendy's, New Balance and Depeche Mode reacted similarly after they were claimed to be the official burgers/shoes/band of the alt-right (Washington Post, 2017; Independent, 2017). In pre-emptive actions Nike and Ben & Jerry's publicly announced that they commit to diversity, in order to prevent that their brands are adopted by hate groups (Washington Post, 2017). Burberry dropped baseball caps from sale and reduced the visibility of their brand pattern to avoid football hooligans as customers (BBC News, 2005). Designers refused to take orders from the finance industry (Bloomberg, 2019) and to dress first Lady Melania Trump (Glamour, 2017). These example demonstrate that firms care about the moral values of their customer pool. Firms also try to avoid customer that have undesirable characteristics unrelated to moral values—although this seems to be less common. Abercrombie & Fitch did not produce XL or XXL sizes in women's clothing to avoid that large women wear their products (Business Insider, 2013) and paid stars from MTV's Jersey Shore to not wear their products (Dunn, White and Dahl 2012). Louis Roederer Champagne made a statement implying that he would prefer it if their products would not be publicly consumed by rappers (Economist, 2006).

Bourdieu, 1984). Because of this identity signaling motive, consumers care about the type-composition of products' customer pools. Models of identity signaling and consumption provide theoretical foundations for brand-image [47], advertising and identity-based consumption (for example, Wernerfelt, 1990; Kuksov, 2007; Vikander, 2017), and have important implications for individual welfare and market outcomes. Identity signaling can result in distortions of consumption expenditures (Frank, 1985; Ireland, 1994), poverty traps (Moav and Neeman, 2010, 2012), Veblen goods (Bagwell and Bernheim, 1996) and fashion cycles (Karni and Schmeidler, 1990; Pesendorfer, 1995).[48] However, the question whether consumers care about the types of others that consume a product as postulated in these models remains open.[49]

In this paper, I provide evidence in favor of consumers caring about the type-composition of products' customer pools. My evidence comes from controlled experimental settings that allow me to manipulate the type-composition of products' customer pools while keeping other aspects of the choice environment constant.[50] I investigate demand for products that do not have properties that connect them directly to specific moral values; potential signals about a consumers characteristics emerge only through the types-composition of products' customer pools.

My work focuses on a specific hypothesis that arises from models of identity signaling. I introduce a simple model of identity signaling and consumption (based on Bernheim, 1994; Bénabou and Tirole, 2011) to make predictions in the environments studied in this paper and to guide my empirical investigation; I do not attempt to provide a substantial novel theoretical contribution. In the model, individuals have imperfect knowledge either of their own or of others' characteristics and care about their (self- or social-) image. I show that consumers conform to the consumption patterns of people with desirable characteristics and avoid products popular among people with characteristics they perceive as undesirable. They do so as a way to assure themselves or others of their desirable

[47] Brand images arise endogenously through the types of people that buy a particular product as part of their equilibrium strategy.
[48] Signaling models have also been proposed to explain how firms set (dynamic) prices (Rao and Schaefer, 2013), product lines (Friedrichsen, 2018) and visibility of products (Carbajal, Hall and Li, 2016).
[49] There is evidence in line with other predictions made by these models of identity signaling, as discussed in Section 2.1.
[50] The hypothesis that consumers care about the type-composition of products' customer pools is difficult to establish or reject in observational data. Changes in the type-composition of customer pools do typically not happen exogenously, but in reaction to events that might change many relevant choice aspects. Moreover, a product's qualities may be tailored to the types of people that consume it, which then might make it less attractive to others.

type.[51] I then test this hypotheses in two laboratory experiments. In the first study, I do so in a public setting in which subjects' social-images are at stake. In the second study, I investigate (non-)conformity in a double-blind setting in which subjects' self-images are at stake.

In the first study, consumption choices are observed by others. I create a situation in which observers know that the customer pool of a product largely consists of either desirable or undesirable types. I vary the desirability of types by treatment. If a subject chooses the product, observers might confuse her with the typical consumer of the product, and attribute the typical consumer's type to her. In this setting, product choices are informative signals about consumers' types. This resembles, for example, the case of the clothing brand Lonsdale in the 1990s and early 2000s; it was public knowledge in Germany that neo-Nazis made up a large share of Lonsdale's customers. A person wearing a Lonsdale sweater in public was likely perceived to have right-wing extremist attitudes. As for types, I use moral values. Moral values might be particularly important in real world applications, as the examples above indicate. In addition to moral values, I also use intelligence to classify people into desirable and undesirable types. Both moral values and intelligence are related to identity and haven been successfully used before to induce image concerns in laboratory experiments (for example, Bursztyn, Egorov and Fiorin, 2017; Zimmermann, forthcoming).

I find that consumers indeed care about the *moral values* of products' customer pools: subjects willingness to pay for a product is statistically significant higher if its customer pool consists of individuals with desirable moral values than if its customer pool consists of individuals with undesirable moral values. The effect size is 14.9% of the average price of the product. Net retail margins for similar products are only about 5% (Damodaran, 2019), so product adoptions by individuals with undesirable characteristics could have sever consequences for the profitability of products and brands. There is no treatment effect for intelligence. This second result demonstrates that there are limits in the extent that customers care about the characteristics of products' customer pools. Taken together, these findings suggest that subjects care more about the public perception of their moral values than about

[51] As a result, brands and products convey identities that shape their consumers' self- and social-images. These "brand-images" arise endogenously through the type of people that choose a particular good as part of their equilibrium strategy (as in Wernerfelt, 1990; Kuksov, 2007; Kuksov, Shachar and Wang, 2013; Kuksov and Wang, 2013, Vikander, 2017).

the public perception of their intelligence.[52] This might explain why most examples for firms avoiding customers come from the domain of moral values.

While consumption often happens in public, many products are also consumed in more private settings than the one investigated in my first study. Papa John's, for example, does home delivery. According to the model, the type-composition of products' customer pools might also matter in private settings due to self-image concerns. In a second study, I investigate consumption in such private settings. I apply a double blind procedure. In the experiment, a product is adopted by right-wing extremists. Conforming to the choices of right-wing extremists might constitute a negative signal about a subject's moral values, in particular racism.

To implement the second study, I collect novel consumption data from right-wing extremists. I recruit 10 neo-Nazis on German right-wing extremist online forums. The neo-Nazis make multiple binary product choices in a short online survey. For the purpose of this study, it is essential that one product is adopted by most neo-Nazis. I achieve this requirement by connecting some products with hidden neo-Nazi symbols, thereby increasing their attractiveness. Some products have, for example, a 88 in their product numbers (a neo-Nazi symbol for HH, "Heil Hitler") while others have the word "milk" (a recent symbol of the alt-right) in their names.

In the laboratory, participants first observe the choices of neo-Nazis in one binary choice situation, and then choose between the same two products. I vary, by treatment, whether neo-Nazis' undesirable identities are revealed to subjects: some subjects are told that they observe neo-Nazis' choices, while others only learn that they observe choices of individuals recruited on the internet. I find that subjects' willingness to pay for the product is lower if it is perceived to be adopted by neo-Nazis then if it is perceived to be adopted by consumers with neutral moral values. The effect size is 9.7% of the price of the adopted product, a large effect in comparisons to retail net margins of 1.7% for similar products (Damodaran, 2019).

In both experiments, I find that subjects care about the moral values of others that consume a product when they make consumption choices. While these findings confirm a key prediction of identity signaling models, they could also be explained by motives unrelated to identity signaling. First, the treatment might have changed the perception of the

[52] This is in line with McManus and Rao (2015), who find that while subjects considered intelligence a desirable trait, they dislike signaling it to others. Unfortunately, I only learned about this study after I implemented my experiment.

products, for example through some form of social learning (e.g., Bikhchandani, Hirshleifer and Welch, 1998). Second, people might be more willing to conform to behaviors of others that have similar moral values than to others that have very different moral values. Third, associating a product with neo-Nazis, as done in Study 2, might result in feelings of disgust when consuming the product—in line with negative contagion (Rozin, Millman and Nemeroff, 1986). Finally, one might be worried about experimenter demand effects, in particular given my use of neo-Nazis as undesirable types in Study 2. My data, including answers to survey questions, allows me to shed some light on subjects' motives. I find that all explanations except identity signaling fail to explain some aspects of the data.

The rest of the paper proceeds as follows. The next section discusses how this paper relates and contributes to the previous literature. Section 3 presents a simple model that demonstrates how identity-signaling can influence consumption choices. Next, I present the design and results from my first and my second study (Section 4 and Section 5, respectively). In Section 6, I discuss potential explanations for my results. I conclude in Section 7 by discussing policy implications of my work.

3.2 Contribution to the literature

My paper relates to several strands of literature, specifically to the literatures on signaling thorough consumption, on identity economics, on conformity and peer effects and on ideology and consumption. In the following, I discuss how my paper relates and contributes to these literatures.

3.2.1 Signaling through consumption

There is a substantial theoretical literature in economics on signaling and consumption, as discussed in the introduction. While most of the earlier work focused on signaling of social-status and wealth (conspicuous consumption), many recent papers take a more general approach and allow consumers to signal other aspects of their identities (Kuksov, 2007; Kuksov, Shachar and Wang, 2013; Kuksov and Wang, 2013; Carbajal, Hall and Li, 2016; Friedrichsen, 2018). The key prediction for consumer behavior in signaling models is that consumers care about the type-composition of products' customer pools.[53]

[53] There are related models that assume that people care about the number (not types) of other consumers that buy a product ("bandwagon effects"; Leibenstein, 1950) and that people care about their relative consumption, that is, how their consumption levels compare to others (e.g., Hopkins and Kornienko, 2004). The latter motive

The empirical literature in economics has focused on wealth signaling through conspicuous consumption of expensive goods (Bloch, Rao and Desai, 2004; Charles et al., 2009; Heffetz, 2011, 2018; Bursztyn et al., 2018; Clingingsmith and Sheremeta, 2018; Cosaerts, 2018) and on signaling of moral values through "moral goods" (Sexton and Sexton, 2014; Delgado, Harriger and Khanna, 2015; Friedrichsen and Engelmann, 2018). None of these studies provide evidence that people care about the type-composition of products' customer pools. The main contribution of this paper to the literature on identity signaling is to provide supportive evidence for this prediction.

My investigation differs in two other important aspects from the earlier empirical literature. First, in the studies cited above, signaling is characterized by a direct link between properties of the product (its price or its "morality") and the image to be signaled (being wealthy or being moral). In my studies, there is no link between properties of the product and a particular identity. Instead, signals only emerge through the types of people that consume particular products. Studying such environments seems important as many products do not exhibit properties that directly relate them to personal characteristics. Second, Study 2 differs from earlier work in that it investigates consumption in private settings. My studies provide evidence that identity signaling extends to private settings and to settings with no link between properties of products and identities, which suggests that identity signaling might be more prevalent in shaping consumer behavior than previously thought.

There is a small literature in marketing that investigates whether consumers care about who else buys a particular product (White and Dahl, 2006, 2007; Berger and Heath, 2007, 2008; and Berger and Rand, 2008). The validity of these studies, however, is limited because they either rely on hypothetical choice situations or suffer from low sample sizes with limited statistical power to detect effects. In addition, it often remains unclear what drives the results.[54] Finally, none of this work investigates conformity to desirable or undesirable types.

3.2.2 Identity Economics

This paper also relates to identity economics. In the model of Akerlof and Kranton (2000), people belong to a particular category (e.g., male or female). Each category is linked

is sometimes also interpreted as status signaling. These models do not make the prediction that consumers care about the type-composition of products' customer pools.

[54] White and Dahl (2006), for example, look at hypothetical choices between steaks of two different sizes. They either call the small steak "chef's cut" or "ladies' cut," and find that men avoid the "ladies' cut." In this setting, treatment differences could be the result of a difference in the perception of whether the small steak satisfies the participant's appetite.

to appropriate behavior (e.g., women should not work as lawyers). If people do not chose an appropriate action, they face costs in terms of a lower self-image. Atkin, Colson-Sihra and Shayo (2019) apply the model of Akerlof and Kranton (2000) to food choices and provide evidence that identity concerns (religion and ethnicity) interact with group-salience and group-status to shape consumption choices, specifically demand for beef and pork in India.[55]

Akerlof and Kranton propose that people's self-images, or identities, can depend on the choices that other types of players make. They write that "a woman working in a 'man's' job may make male colleagues feel less like 'men'," which then makes the job less attractive for men.[56] My studies provide support for such effects in the domains of consumption and moral values.

3.2.3 Conformity and peer effects

My investigation also relates to the literature on conformity and peer effects (e.g., Krupka and Weber, 2009; Zafar, 2011; Bernheim and Exley, 2015; Lahno and Serra-Garcia, 2015; Gioia, 2017; Bigenho and Martinez, 2018). Bernheim (1994) and Andreoni and Bernheim (2009) study how image concerns can result in conformity to pro-social behavior. Fatas, Hargreaves Heap and Rojo Arjona (2018) and Dimant (2019) present some evidence that conformity to pro-social behavior can depend on similarity to peers. I add to this literature by studying whether people's conformity decisions depend on the desirability of others' types.

3.2.4 Ideology and consumption

Groups with different political ideologies consume different products and brands (e.g., Gebru, et al. 2017; Bertrand and Kamenica, 2018; Kapner and Chinni, 2019).[57] A recent literature in political sciences, psychology and marketing investigates why ideology predicts consumption, focusing on differences in tastes and personality (Khan, Misra and Singh, 2013; Kidwell, Farmer and Hardesty, 2013; Roos and Shachar, 2014) and on firms' moral values (McConnell, et al., 2018). My study contributes to this literature by demonstrating that consumers care about the moral values (or, ideology) of products'

[55] Atkin, Colson-Sihra and Shayo (2019) assume a exogenously given connection between types and appropriate consumption patterns. Signaling models differs in that "appropriate behavior" for different types of players arise endogenously through the type of people that choose a particular product.

[56] Akerlof and Kranton model this intuition by allowing the utility of the male decision maker to depend on the choice of the female worker. Models of identity signaling give a micro-foundation for such an assumption: an increase in the share of female workers in a job weakens the signal of workers' masculinity, thereby threatening male workers' self-images and decreasing their utility.

[57] Cambridge Analytica even used fashion preferences to identify right-wing voters during 2016 presidential election (New York Times, 2018).

customer pools. Differences in consumption decisions between conservatives and liberals might be partly explained by consumers' desire to signal their political positions, and to differentiate themselves from people with the opposite ideology.

3.3 A simple model of consumption and identity signaling

In this section, I introduce a simple model that illustrates how consumers can signal desirable characteristics through consumption choices. I build on the frameworks of Bernheim (1994) and Bénabou and Tirole (2011). The key prediction of the model is that customers care about the type-composition of products' customer pools. I do not attempt to provide a novel theoretical contribution, but rather use simple economic analysis to make predictions in the environments studied in this paper and to guide my empirical investigation.

Consider a population of consumers, normalized to [0,1], each of whom chooses between two products A and B. Each consumer *i* has a type $t \in \{A, A_c, B\}$. While A- and A_c-types receive more consumption utility, u^t, from product A, B-types derive more consumption utility from product B: $u^A(A) - u^A(B) = u^{A_c}(A) - u^{A_c}(B) = u^B(B) - u^B(A) = \Delta u > 0$. The distribution of types is common knowledge and described by $\Pr(A) = \delta$, $\Pr(A_c) = \gamma$ and $\Pr(B) = 1 - \gamma - \delta$, with $\delta, \gamma > 0$ and $\delta + \gamma < 1$. Consumers' types are private information.

Consumers have some characteristics, for example moral values, that shape their (self- or social-) image, v^t, with $v^A = v^B = v^{-c} > 0$. The focus of this analysis are the characteristics of the A_c-types, c. In the following, I will compare a situation in which the A_c-types have desirable characteristics, c=g, to the situation in which the A_c-types have undesirable characteristics, c=b. In the former case v^c corresponds to $v^g > v^{-c}$, while in the latter case v^c corresponds to $v^b < v^{-c}$. I will demonstrate that if consumers care about their image, the choices of the A- and B-types depend on the desirability of the A_c-types' characteristics: the product A is more popular among the A- and B-types if the A_c-types have desirable characteristics than if the A_c-types have undesirable characteristics.

Consumers care about the public perception of their type (that is, their social-image), $E(v|x)$. A consumer's social-image is calculated using Bayes' rule and depends on the consumer's choice, the choices of all other consumers, and the distribution of types. The probability that a consumer has characteristics c given her choice $x \in$ {A,B} is given by:

$$\rho(A) = \frac{\gamma x^{A_c}}{\delta x^A + (1-\gamma-\delta)x^B + \gamma x^{A_c}}$$

$$\rho(B) = \frac{\gamma(1-x^{A_c})}{1-\delta x^A-(1-\gamma-\delta)x^B - \gamma x^{A_c}}$$

where x^t is the probability that a consumer with type *t* chooses product A. A consumer's social-image is then given by $E(v|x) = \rho(x)v^c + \big(1-\rho(x)\big)v^{-c}$. A- and B-type consumers choose $x \in \{A, B\}$ that maximize the following utility function:

$$u^t(x) + \alpha E(v|x),$$

where $\alpha > 0$ is the weight they put on their social-image. A_c-types do not care about their social-image, that is $\alpha^{A_c} = 0,$ which implies that $x^{A_c} = 1$.[58]

This model can also be interpreted as a self-signaling model, following Bénabou and Tirole (2011). For this alternative interpretation, players have imperfect knowledge of their own types. Consumers' past choices are then signals about their own type. There are two periods. In period 0, the consumer "obtains a momentary insight into his true nature," (Bénabou and Tirole, 2011) and temporarily learns his type. Then, he makes his choice. In period 1, the consumer remembers his type t with probability $1-\alpha \in (0,1)$, with probability α the consumer has no direct access to the motivation behind his behavior in period 0 (that is, his type), and only remembers his choice, *x*. In period 0, the consumer cares about his consumption utility and his expected self-image in period 1: he maximizes $u^t(x) + (1-\alpha)v^t + \alpha E(v|x)$. Note that this consumer problem corresponds to the consumer problem discussed before.

In the following, I will study how the consumption decisions of the A- and B-types depend on the characteristics of the A_c-types. Remember that the A_c-types choose A. I investigate how the share of A- and B-types that choose A, $x^{A\cup B} = \frac{\delta x^A + (1-\gamma-\delta)x^B}{1-\gamma}$, depends on *c*. I refer to these shares as $x_g^{A\cup B}$ and $x_b^{A\cup B}$ for *c*=*g* and *c*=*b*, respectively. I follow Bénabou and Tirole (2011) by restricting attention to monotonic Perfect Bayesian equilibria. In a monotonic equilibrium, $x^A = x^B = x^{A_c} = 1$ implies $\rho(B) = 0$. I characterize all monotonic equilibria in Appendix G. If *c* is a desirable characteristic (*c*=*g*), there is a unique monotonic equilibrium. If *c* is an undesirable characteristic (*c*=*b*), then there might exist up to three equilibria.

[58] An alternative assumption with similar implications is to assume the A_c-types receive substantially more consumption utility from product A than from product B: $u^{A_c}(A) - u^{A_c}(B) > \alpha * max(v^g - v^{-c}, v^{-c} - v^b) > u^A(A) - u^A(B) = u^B(B) - u^B(A) = \Delta u$.

The Proposition shows that, as long as the difference in consumption utility between the two goods (Δu) is not too large, more consumers choose product A if it is popular among people with desirable characteristics than if it is popular among people with undesirable characteristics. Given that there can be multiple equilibria for *c=b*, the Proposition shows that $x_g^{A \cup B}$ is (weakly) bigger than $x_b^{A \cup B}$ for *all* possible equilibria. To do so, I compare $x_g^{A \cup B}$ with the equilibrium for *c=b* with the highest share of consumers that choose A, $\max(x_b^{A \cup B})$.

The intuition behind the Proposition is that if product A is popular among desirable types, then choosing product A increases the consumer's image. This makes the product relatively more attractive than product B, and even B-types might choose product A. If product A is popular among consumers with undesirable types, then choosing product A decreases the consumer's image, and even the A-types might avoid it. The Proposition also shows that the difference between the *c=g* and the *c=b* context is more pronounced when subjects care more about their image, that is, if α is higher.

Proposition. The relationship between $x_g^{A \cup B}$ and $x_b^{A \cup B}$ is characterized by a threshold $\overline{\Delta u}$ such that $x_g^{A \cup B} > \max(x_b^{A \cup B})$ if $\Delta u < \overline{\Delta u}$ and $x_g^{A \cup B} = \max(x_b^{A \cup B}) = \frac{\delta}{1-\gamma}$ if $\Delta u \geq \overline{\Delta u}$. Furthermore, $\overline{\Delta u}$ is increasing in α.
Proof: See Appendix G.

The model demonstrates that consumers care about the type-composition of a product's customer pool, but only if α is high enough (in relation to Δu). In my laboratory experiments, I use characteristics *c* that are likely seen as very undesirable or very desirable byß many participants, and therefore should result in a high α. Study 2 looks at consumption choices in private settings. In such settings, consumption might serve as self-signals about consumers' types. According to the model, the key requirement for self-signaling is that participants are insecure about their own types (that is, α is not 0). As I will discuss in more detail later, I attempt to increase participants' insecurity with means of an identity threat.

3.4 Study 1: Consumption in public settings

In Study 1, I investigate consumption in a setting where choices are observed by others. I create a situation in the laboratory in which it is public knowledge that the customer

pool of one product largely consists of people with either desirable or undesirable characteristic. Consumption choices are then informative signals about subjects' characteristics; if a subject choses the product that is adopted by the desirable or undesirable types, she might be confused with the typical consumer of the product.[59] To test whether consumers care about the type-composition of product's customer pools, I vary, by treatment, whether the customer pool consists of people with desirable or with undesirable characteristics.

3.4.1 Experimental design

In the laboratory, I first elicit measures of participants' intelligence and moral values. Next, I investigate how individuals' consumption decisions respond to the potential that their choices signal information about either their intelligence or their moral values to observers. Depending on a subject's consumption choice, observers might confuse the participant with another subject with desirable (*desirable treatment*) or undesirable characteristics (*undesirable treatment*).

To decide on multiple aspects of the design, including which measure of moral values and which products to choose, I implemented a short online survey.[60] I recruited 29 participants drawn from the same subject pool from which I recruit participants for my laboratory experiment. I will refer to this sample as the *online survey*.

3.4.1.1 Measuring intelligence

I measure participants' intelligence with a test consisting of 12 Raven's matrices. Raven's matrices have been successfully used in other studies to induce image concerns (for example, Zimmermann, forthcoming). Subjects see patterns in which one part is missing. For each pattern, they are given 8 possible suggestions for how to complete it. Subjects have twelve minutes time to complete the 12 patterns. For every correct pattern, they earn CHF 0.50 (CHF 1 ≈ USD 1). Before subjects start with the test, they solve two training patterns. Subjects are told that Raven's matrices are regularly used to measure general intelligence. They do not receive feedback about their performance in the test.

[59] This resembles, for example, the case of the clothing brand Lonsdale in the 1990s and early 2000s; it was well known in Germany that neo-Nazis made up a large share of Lonsdale customers. A person that was wearing a Lonsdale sweater in public was likely perceived to have right-wing extremist attitudes.

[60] In the online survey, subjects repeatedly choose between pairs of products. I measure subjects' preferences in the same way is described in paragraph 3.4.1.3. Next, subjects made choices in four choice situations designed to measure their moral values, including the one used in the laboratory (see paragraph 3.4.1.2). In all choice situations, people face tradeoffs between their own payoff, and a donation to an organization. Finally, subjects have to rate whether they like or dislike it if they would be publicly associated with these organizations.

3.4.1.2 Measuring moral values

To measure moral values, I give subjects the option to increase their payoff by authorizing the researchers to make a donation to Zukunft CH, a conservative Christian organization, on their behalf. They are told that the members of Zukunft CH fight "the sneaking introduction of the sharia" and marriage for same-sex couples, engage in demonstrations against abortions and advocate conversion therapies for homosexuals. Individuals dislike being publicly associated with Zukunft CH: participants in the online survey report that they would dislike it to be perceived as a donor of this organization (for details, see Figure F1 in the Appendix).

Subjects receive an endowment of 6 CHF. Then, they have to choose one of the seven options shown in Table 8, each links a payment to the subject with a donation to Zukunft CH.

Table 8. Choice situation to measure moral values

Options	In addition to the CHF 6.00 endowment, the subject receives	Donation to Zukunft CH on behalf of subject
Option 1	+6.00 CHF	+9.00 CHF
Option 2	+4.00 CHF	+6.00 CHF
Option 3	+2.00 CHF	+3.00 CHF
Option 4	+0.00 CHF	+0.00 CHF
Option 5	-2.00 CHF	-3.00 CHF
Option 6	-4.00 CHF	-6.00 CHF
Option 7	-6.00 CHF	-9.00 CHF

I include both positive and negative donations to increase the perceived discrepancy between the "morally good" and the "morally bad" actions, and to avoid that most subjects choose options on the boundary (Option 1 or Option 7). Negative donations are explained as "preventing donations from other participants from being implemented." To implement subjects' choices, all individual donations are added up.[61] Donations are anonymous and are not subtracted from the donors' payments. Subjects are told that other study participants might receive some (incomplete) information about their choices.

3.4.1.3 Consumers' choices

Next, subjects choose between two products. There are two rounds of product choice. Rounds differ in the choice sets subjects face. In one round, subjects choose between two

[61] Note that this procedure does not preclude a negative total donation. Data from the subjects in the online survey suggested that this is not an issue. Indeed, the final experiment resulted in a total donation of CHF 237 to Zukunft CH.

packs of chocolate bars, one of them is produced by Camille Bloch and the other by Munz. Both packs consist of 5 chocolate bars, 23 grams each, and are priced at about CHF 3.50. In the other round, subjects choose between a cup and a 4GB USB stick. The market price of the cup and the 4GB USB stick is about CHF 8 and CHF 4, respectively.[62] I randomize, at the session level, which choice set subjects face first.

Instead of one binary choice, I elicit subjects' willingness to pay to receive one product instead of the other product. To do so, participants make 13 decisions between bundles of products and money, as shown in Table 9 for the Munz and Camille Bloch chocolate. At the end of the experiment, one of the thirteen cases is randomly drawn for each round and might be implemented.

Table 9. Consumption choices

Decision	Choice situation		
1	Munz + 3.00 CHF	*or*	Camille Bloch
2	Munz + 2.50 CHF	*or*	Camille Bloch
3	Munz + 2.00 CHF	*or*	Camille Bloch
4	Munz + 1.50 CHF	*or*	Camille Bloch
5	Munz + 1.00 CHF	*or*	Camille Bloch
6	Munz + 0.50 CHF	*or*	Camille Bloch
7	Munz	*or*	Camille Bloch
8	Munz	*or*	Camille Bloch + 0.50 CHF
9	Munz	*or*	Camille Bloch + 1.00 CHF
10	Munz	*or*	Camille Bloch + 1.50 CHF
11	Munz	*or*	Camille Bloch + 2.00 CHF
12	Munz	*or*	Camille Bloch + 2.50 CHF
13	Munz	*or*	Camille Bloch + 3.00 CHF

3.4.1.4 Composition of products' customer pool

The following describes how I manipulates the type-composition of products' customer pools, the key feature of my design. Participants' consumption choices are revealed to a set of *observers*. To induce social-image concerns, their choices are linked to portrait pictures.[63] In addition to information about the consumption choices of the participant (the *consumer*), observers also receive information about the consumption choice of another

[62] I selected the 4GB USB stick and the cup, and the two chocolates, because the responses in the online survey suggested that the distribution of the willingness to pay to receive the 4GB USB stick instead of the cup, and the Munz chocolate instead the Camille Bloch chocolate is symmetric with a mean of zero. (However, as I will discuss later, in my actual study most subjects preferred the stick over the cup.)
[63] The picture is taken at the very beginning of the experiment, before subjects enter the lab.

subject, the *target*. The target is a participant that is selected due to his characteristics, either due to his intelligence-score or due to his moral values. Based on the consumer's choices, the observers might confuse the consumer with the target, and attribute the target's characteristics to her. I vary, by treatment, whether the target's characteristics are desirable or undesirable.

Targets and consumers: For each round of consumption choice, two participants are selected to play the role of the targets in this round. Targets differ between rounds. In one round, targets are selected due to their intelligence score. The subject with the lowest intelligence score among all subjects in the session is selected to be the target with undesirable characteristics and the subject with the highest intelligence score is selected to be the targets with the desirable characteristics.[64] In the other round, targets are selected due to their donation to Zukunft CH. The subject that made the highest donation to Zukunft CH is selected to be the target with undesirable characteristics and the subject that made the lowest donation to Zukunft CH is selected to be the target with the desirable characteristics. If multiple subjects qualify to play the role of a target, one subject is randomly selected.

If a participant is not a target in a given round, she plays the role of a consumer in this round. Therefore, in a session with *N* participants, each round consists of two targets, and *N-2* consumers.

Treatments: In each round, first the round's targets choose between the round's products. Targets make only a binary choice between the two products, none of them is bundled with money. Next, the consumers of this round learn the type and choice of one target. Which target they observe depends on the treatment:

- Subjects in the *undesirable treatment* learn the type and choice of the target with undesirable characteristics.
- Subjects in the *desirable treatment* learn the type and choice of the target with desirable characteristics

The treatment is randomized within session. Subjects are assigned to the same treatment for both rounds. After consumers learned the type and choice of their target, they choose between the round's products in the 13 cases, as described in paragraph 3.4.1.3.

Suppose, for example, that the choice set in the first round is the two kinds of chocolates, and the targets are selected due to their intelligence score. The subject with the lowest intelligence score chose the Munz chocolate. Consumers in the undesirable treatment

[64] To avoid negative effects on subjects' self-images, the targets do not learn that the had the lowest, or highest intelligence score. They only learn that others might observe their choices.

are then told that the "participant with the lowest intelligence score chose the Munz chocolate bars," and are then called upon making their own choices.

Observers: At the end of the study, consumers' choices and pictures are seen by up to 14 observers.[65] Depending on a consumer's choices, the observers might confuse the consumer with her target. The observers see either the product choice and picture of the consumer, or the product choice and picture of the consumer's target, depending on the flip of a computerized fair coin. Importantly, observers do not learn the outcome of the coin flip and therefore do not directly learn whether they see the consumer or the target. However, observers are told the type and choice of the consumer's target. Based on the choice of the target and the choice of the observed participant, the observers can draw conclusions about the likelihood that the observed participant is the target.

Remember that the consumer makes choices for 13 cases. If the coin flip selects the consumer, the observer does not see the choices for all 13 cases, but he only sees which product the consumer chose for one randomly drawn case, the *drawn case*. He is not told the amount of money that was bundled with the two products for the drawn case.

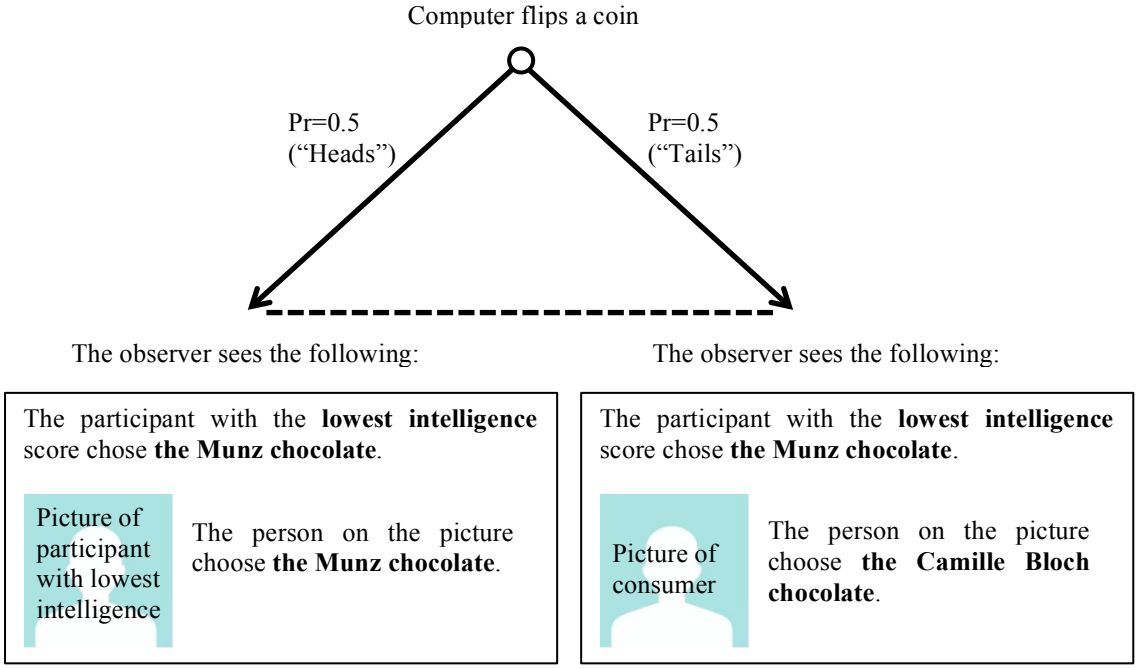
Figure 10 illustrates this procedure with an example. In the example, the consumer is paired with the target with the lowest intelligence score. The target chose the Munz chocolate. The consumer chose the Camille Bloch chocolate for the drawn case. Independent of the coin flip, the observer learns the type and the choice of the consumer's target ("The participant with the lowest intelligence score chose the Munz chocolate."). If heads is drawn, the observer sees to target's picture and product choice. If tails is drawn, the observer sees the consumer's picture and product choice. In the latter case, the observer can conclude that the person on the picture is the consumer, given that her choice differs from the target's choice. If the consumer would have conformed to her target, the observer could not tell whether she is the consumer or the target.

As a result of this procedure, observers should think that the customer pool of one product consists to a large share ($\geq 50\%$) of customers with either desirable or undesirable characteristics. Figure H1 in the Appendix gives the observers' instructions.

Consumers learn all of this before they make their consumption choices.

[65] The role of observers are played by other consumers. However, this is not announced. The exact number of observers depends on the size of the specific session and is in between 11 and 14 observers. As I discuss in the next paragraph, this number was chosen as part of the procedure to guarantee targets' anonymity.

Figure 10. Illustration observer information



Note: The observer does not know whether tails or heads realized, and therefore does not know whether he observes the picture and choice of the consumer or of the target. This is illustrated by the dotted line. Note, however, if tails is drawn, he should conclude that the drawn participant is the consumer.

Protecting targets' privacy: To protect targets' privacy, I implement a procedure that guarantees that targets are never revealed with certainty to observers. As a result of this procedure, each round only counts for 3 consumers in a session. For these 3 consumers, one of the 13 cases is randomly drawn to be implemented (the drawn case), and is potentially seen by two thirds of all consumers, depending on the coin flips.[66] The choices of all other consumers are neither implemented nor seen by observers. Targets' choices are always implemented.

[66] For each round of product choice, each consumer plays the role of an observer for two other consumers. These two consumers might share the same target. If an observer sees the same picture twice, it would imply that the observed person is the target. The following procedure guarantees that no observer will see the same participant twice by introducing a correlation between the two coin flips: observers' two coin flips are either tails-heads or heads-tails (but never heads-heads or tails-tails). Starting with round 1, all consumers of this round are randomly assigned to three groups of equal size (groups 1 – 3). Then, I randomly draw one member of each group. The choices of these three subjects count. The members of group 1 play the roles of observers for group 2 and group 3. Each member of group 1 either observes the drawn member of group 2 (=tails) and the target of the drawn member of group 3 (=heads), or the target of the drawn member of group 2 (=heads) and the drawn member of group 3 (=tails). In a similar manner, the members of group 2 play the role of observers for group 1 and group 3, and the members of group 3 play the role of observers for group 1 and group 2. The same procedure is then repeated for the second round. This procedure corresponds, from a subject's perspective, to the procedure explained above. Note that each drawn member is observed by 2/3 of the consumers, resulting in between 11 and 14 observers.

Sequence: After participants' intelligence and moral values are measured, the first round starts. The first rounds' targets make their binary choices. At the same time, the first rounds' consumers receive instructions that explain in detail all aspects of the first round, including their target's type and that their choices will be seen by the observers. After consumers read the instructions, they answer understanding questions and then learn their targets' choices. Next, they make their choices for the 13 cases. Before choices are revealed to observers, subjects do the second round. The second round comes as a surprise; subjects only know in advance that there will be another part in the experiment. The second rounds' targets then make their choices, while the second round's consumers read the instructions. Then, consumers learn their targets' choices, and make their own choices for the 13 cases. Finally, consumers' choices for both rounds are revealed to observers.

3.4.1.5 Questionnaire

At the end of the experiment, but before consumers' choices are revealed to observers, subjects fill out a short questionnaire. In the questionnaire, I measure subjects' perceptions of the four products used in the experiment. In addition, I elicit subjects' willingness to pay to receive each of the products in private, and I measure how much subjects care about being perceived as intelligent and tolerant.

3.4.1.6 Procedure

I conducted eight sessions, each consisting of between 19 and 24 participants. In total, 170 subjects participated in the study, with 87 participants in the desirable treatment and 81 participants in the undesirable treatment.[67] All sessions took place at the Laboratory for experimental and behavioral economics at the University of Zurich, in June 2019. Participants were recruited using hroot (Bock, Baetge and Nicklisch, 2014) from the joint subject pool of the University of Zurich and the Swiss Federal Institute of Technology (ETH). The experiment was implemented using z-Tree (Fischbacher, 2007). The Online Appendix supplies the instructions for the study.

3.4.2 Results

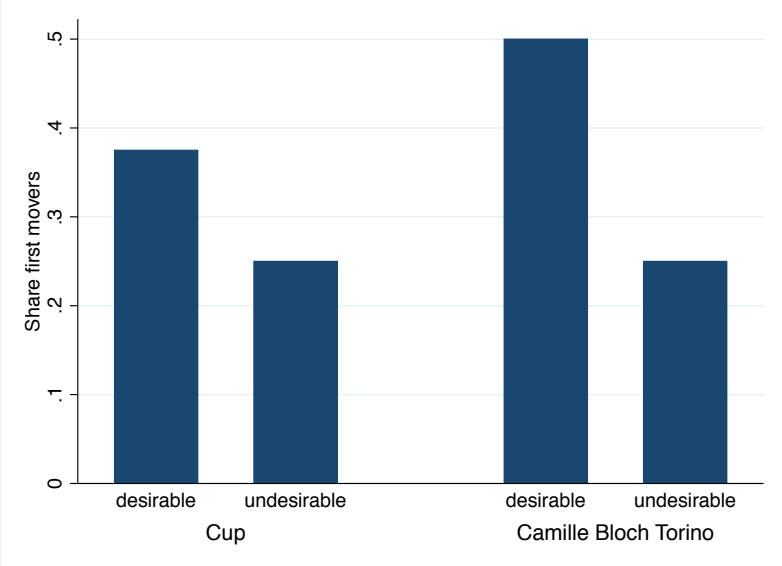
On average, subjects solved 8.41 out of 12 Raven's matrices (std. dev. = 2.01). Only 4.71% of subjects solved all 12 matrices, and the minimal number of matrices solved was 3. For moral values, the median (and modal) choice was the option that neither increased the

[67] Two subjects were in the role of targets in both rounds. For these subjects, I do not have any observations about their behavior in the role of consumers and, as a result, they are not allocated to any of the two treatments.

subject's payment nor the donation to Zukunft CH (Option 4). 32.35% of subjects chose the option that maximized their own payoff and the donation to Zukunft CH (Option 1), while 14.12% of subjects chose the option that minimized the donation to Zukunft CH and their own payoff (Option 7). Figure F2 and F3 in the Appendix provides the full distribution for moral values and intelligence, and illustrates how the distributions of targets compares to the distributions of the consumers.

Regarding targets' behavior, 6 out of 16 targets chose Camille Bloch Torino chocolate over Munz chocolate, and 5 out of 16 targets chose the cup over the 4GB USB stick. Figure 11 shows that the targets' choices are unbalanced between treatments.[68] I will account for these differences in the analysis of consumer behavior.

Figure 11. Targets' choices in the two treatments



Notes: "Cup" gives the share of targets in each treatment that chose the cup over the USB stick. "Camille Bloch Torino" gives the share of targets in each treatment that chose the Camille Bloch Torino chocolate over the Munz chocolate.

In the following, I will focus on consumers' behavior. Remember that consumers chose between bundles of products and money. Six subjects made product choices that are non-monotone in money,[69] one subject did so in both rounds. These seven observations are

[68] Neither the treatment difference in the share of targets that chose the cup nor the treatment difference in the share of targets that chose the Camille Bloch Torino chocolate is statistically significant different from zero (tests of proportions, z=0.54 (p=0.59) and z=1.03 (p=0.30), respectively).

[69] They chose (A, CHF X) over (B, CHF 0) but (B, CHF 0) over (A, CHF X-0.5), where A and B are either the two kinds of chocolates or the cup and the USB stick.

excluded from my analysis.[70] The fact that most subjects exhibit monotone choice patterns can be seen as an indicator that subjects understood the choices they were facing.

I will now turn to the question whether the targets' characteristics affect the choices of the consumers, the focus of this paper. I start my analysis of consumer behavior by looking at the binary choice between the two products when none of them are bundled with a monetary payment; do subjects pick the option that was chosen by their target less often in the undesirable treatment than in the desirable treatment?[71] Table 10, column (1) gives the estimated coefficients of a linear regression of the probability of choosing the same option as the target on a treatment dummy (1 = undesirable treatment). Column (2) adds session fixed effects to the specification. I find support for the hypothesis that consumers care about the type-compositions of products' customer pools: a subject is estimated to be 16.8 percentage points less likely to chose the same product as her target in the undesirable treatment than in the desirable treatment (p=0.003).

Remember that targets' choices are unbalanced between treatments. This can introduce a bias in the estimation of the treatment effect.[72] The first two specification likely underestimate the treatment effect because most subjects preferred the USB stick over the cup[73] and more targets chose the USB stick in the undesirable treatment than in the desirable treatment. The specification in column (3) accounts for this issue by controlling for the targets' choices. In specification (3), the estimated treatment effect increases to 20.9 percentage points. Alternatively, this issue can be addressed by estimating treatment effects separately for each possible target choice. Figure F4 (a) in the Appendix shows that results are qualitatively similar under this approach.

[70] Table F1 and F2 in the Appendix shows that results do not change if I keep these observations.
[71] I preregistered participants' willingness to pay to receive the same product as their targets instead of receiving the other product as the main outcome variable (AEARCTR-0004268). I will discuss the willingness to pay next. Note that, while results are qualitatively similar for both outcome variables, there is a difference in terms of significance (see Table 10).
[72] Here is an example to illustrate how this can result in a bias. Suppose that 66% of subjects prefer the USB stick over the cup, and that there is no treatment effect. However, suppose that by chance 80% of targets in the desirable treatment choose the cup while 80% of the targets in the undesirable treatment choose the USB stick. The expected number of subject that choose the same product as their target is 40.4% in the desirable treatment and 59.6% in the undesirable treatment. The treatment difference is therefore (wrongly) estimated to be 19.2 percentage points. If controls for the product choice of the target are added, however, the treatment effect is estimated correctly at zero.
[73] 64.7 percent of subjects chose the cup over the USB stick when none of them are bundled with a monetary payment. In the survey, subjects report a on average CHF 0.43 (t=-2.56, p=0.011) higher willingness to pay for the USB stick than for the cup in private. Also, 67.6 percent of subjects that have a strict preference between the two products have a higher willingness to pay for the USB stick than for the cup.

Table 10. Treatment effects

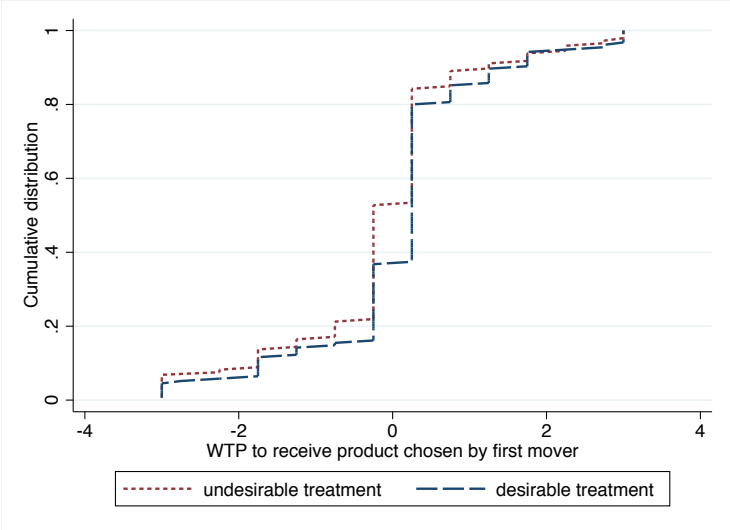
Dependent variable:	Pr(Conform to target)			WTP for target's product		
	(1)	(2)	(3)	(4)	(5)	(6)
1 = undesirable treatment	-0.160*** (-2.85)	-0.168*** (-3.01)	-0.209*** (-3.80)	-0.229 (1.50)	-0.267* (-1.82)	-0.329** (-2.20)
1 = target chose Munz			0.166* (1.79)			0.160 (0.78)
1 = target chose cup			-0.045 (-0.44)			-0.193 (-0.63)
1 = target chose USB stick			0.240*** (2.98)			0.400* (1.91)
Constant	0.632*** (16.34)			-0.082 (0.78)		
Log(sigma)				1.337*** (15.06)	1.310*** (15.00)	
N	301	301	301	301	301	301
Session FE	No	Yes	Yes	No	Yes	Yes

*Notes: (1)-(3): Linear regressions of probability to choose the same product as the target when none of the two products are connected with any payment on a treatment dummy, a constant and, depending on the specification, session fixed effects and controls for the targets' choices. (4)-(6): Tobit regressions (left-censored at CHF -3, n=17; right-censored at CHF +3, n=10) of willingness to pay to receive the same product as the target instead of the other product on the same set of variables. t-statistics in parentheses; standard errors are clustered at subject level (168 clusters); * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

The choice data collected allows me to calculate participants' willingness to pay to receive the same product as their targets instead of receiving the other product.[74] This is the main outcome variable of this analysis. Note that it is left-censored at CHF -3 (10 observations) and right-censored at CHF 3 (17 observations). Figure 12 shows the cumulative distributions of the willingness to pay in both treatments. In both treatments, the willingness to pay for most subjects is in between CHF -1 and CHF 1.

[74] Note that my data is discrete due to use of the list method. Hence, for a subject, who choses Camille Bloch in the choice situation "(Camille Bloch, CHF 1.50) or (Munz, CHF 0)", but choses Munz in the choice situation "(Camille Bloch, CHF 1) or (Munz, CHF 0)", the difference in monetary value must be in [CHF 1, CHF 1.5]. I set this difference to CHF 1.25.

Figure 12. Distribution willingness to pay



Notes: Cumulative distributions of subjects willingness to pay to receive the product chosen by the target instead of the other product. (Figure F5 in the Appendix shows the distributions of willingness to pay after controlling for session fixed effects and targets' choices.)

Table 10, column (4) gives the estimates from a Tobit regression of the willingness to pay on a treatment dummy. The specifications in column (5) adds fixed effects, and the specification in column (6) controls for targets' choices. The estimates supports my hypothesis in that the coefficient are negative in all three specifications. That is, consumers willingness to pay is lower for a product that is popular among people with undesirable characteristics than for a product that is popular among people with desirable characteristics. However, the effect is only significant (at the 5% level) in the specification that controls for targets' choices. Results are qualitatively similar (and statistically significant) if treatment effects are estimated separately for each possible target choice (see Figure F4 (b) in the Appendix).

Next, I will investigate whether treatment effects differ for moral values and intelligence. Table 11 replicates the previous analysis, but estimates treatment effects separately for moral values (TE-M) and intelligence (TE-I). While there is a significant treatment effect for moral values in all six specifications, none of the specifications yields a significant treatment effect for intelligence. For the willingness to pay measure, I can reject the hypothesis that there is no difference between treatment effects for moral values and intelligence. Results are qualitatively similar if treatment effects are estimated separately for the target choices (see Figure F4, (c) – (f) in the Appendix).[75]

[75] I also pre-registered that I will look at potential differences between choice sets. Participants' decisions do not depends on which pair of objects was used (the two kinds of chocolates or the cup and the USB stick). If I

Table 11. Treatment effects for intelligence and moral values

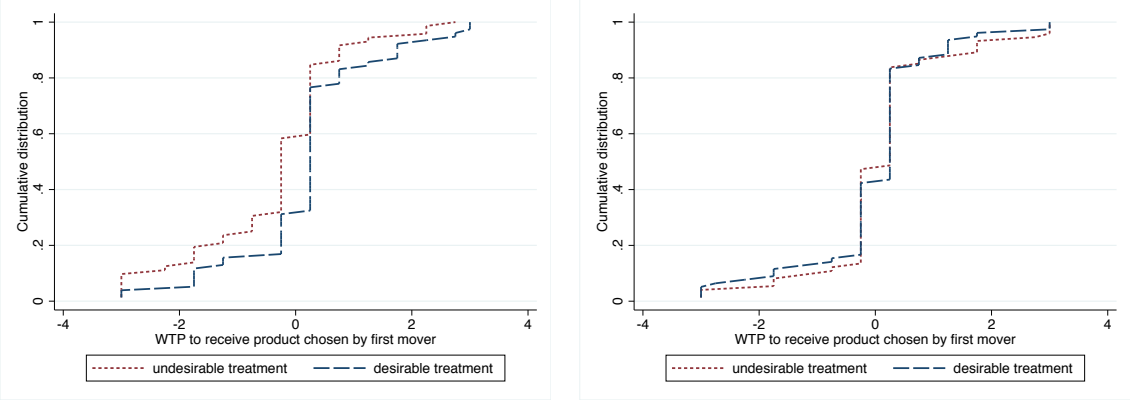
Dependent variable:	**Pr(Conform to target)**			**WTP for target's product**		
	(1)	(2)	(3)	(4)	(5)	(6)
1 = undesirable treatment	-0.272***	-0.277***	-0.290***	-0.600***	-0.623***	-0.652***
* 1 = moral values (TE-M)	(-3.43)	(-3.52)	(-3.76)	(-2.65)	(-2.86)	(-3.04)
1 = undesirable treatment	-0.050	-0.109	-0.128	0.134	0.082	-0.009
* 1 = intelligence (TE-I)	(-0.61)	(-1.40)	(-1.56)	(0.64)	(0.39)	(-0.04)
1 = intelligence round	-0.111	-0.109	-0.015	-0.209	-0.197	-0.010
	(-1.43)	(-1.40)	(-0.18)	(-0.97)	(-0.92)	(-0.05)
Constant	0.688			0.187		
	(12.94)			(1.22)		
Log(sigma)				1.322***	1.296***	1.277***
				(15.25)	(15.18)	(14.85)
p-value TE-M==TE-I	0.055	0.065	0.161	0.020	0.024	0.040
N	301	301	301	301	301	301
Session FE	No	Yes	Yes	No	Yes	Yes
Target choice controls	No	No	Yes	No	No	Yes

*Notes: (1)-(3): Linear regressions of probability to choose the same product as the target when none of the two products are connected with any payment on the interaction between a treatment dummy and a dummy for being in the moral values round (TE-M), the interaction between a treatment dummy and a dummy for being in the intelligence round (TE-I), a dummy for being in the intelligence round, (1=intelligence round), a constant and, depending on the specification, session fixed effects and controls for the targets' choices. (4)-(6): Tobit regressions (left-censored at CHF -3, n=17; right-censored at CHF +3, n=10) of willingness to pay to receive the same product as the target instead of the other product on the same set of independent variables. t-statistics in parentheses; standard errors are clustered at subject level (168 clusters); * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Figure 13 gives the cumulative distribution functions of subjects willingness to pay conditional on treatment and round (intelligence or moral values). The figure confirms the results in Table 11. For moral values, the treatment shifted the entire distribution, while there is no such effect for intelligence.

replicate Table 11 with 1=pair of chocolates and the corresponding interactions, I do not find any evidence for differences in treatment effects between the chocolate round and the cup and UBS stick round (p-values>0.80 for all specifications).

Figure 13. Distribution willingness to pay for moral values and intelligence



(a) moral values (b) intelligence

Notes: Cumulative distributions of subjects willingness to pay to receive the product chosen by the target instead of the other product. (a) use only data from the moral values round and (b) uses only data from the intelligence round. (Figure F5 in the Appendix shows the distributions of willingness to pay after controlling for session fixed effects and targets' choices.)

These findings suggest that subjects care more about the public perception of their moral values than the public perception of their intelligence. In terms of the model, this means that $\alpha_{moral\ values}$ is bigger than $\alpha_{intelligence}$. This is in line with McManus and Rao (2015), who find that while subjects considered intelligence a desirable trait, they dislike signaling it to others.

How should the effect sizes found in this study be interpreted? The average price of the product adopted by the target is CHF 4.39.[76] In the pooled data, subjects' willingness to pay is estimated to be CHF 0.329 lower if the product is adopted by the targets with undesirable identities—a decrease of 7.5% of the average product price. For moral values, the treatment effect is even estimated to be CHF -0.652—a decrease of 14.8% of the average product price. Given that European net margins for food processing, food retail, general retail, and electronics are only 6.76%, 1.67%, -0.48% and 11.83%, respectively (Damodaran, 2019),[77] product adoptions by customers with undesirable characteristics could have sever consequences for the profitability of products and brands.

This first study investigates consumption in a setting where choices are observed by others. While it is important to study such settings because consumption often happens in

[76] The Camille Bloch chocolate, the Munz chocolate, the cup and the USB stick served as the target's choice for 62 consumers, 89 consumers, 48 consumers and 102 consumers, respectively. The average price of the product adopted by the target is therefore 62/301*3.5+89/301*3.5+48/301*8+102/301*4=4.39.

[77] To calculate an average net retail margins for products similar to the ones used in this study, I use the food retail margins (1.67%) for Camille Bloch and Munz, general retail margins (-0.48%) for the cup and electronics margins (11.83%) for the USB stick. These margins are then weighted by the frequency in which each product served as the target's choice for a consumer. The resulting average net margin is 4.8%.

public and has the potential to signal characteristics to others, many products are also consumed in more private settings. Do the moral values of a product's customer pool also matter in such contexts? I address this question in a second study.

3.5 Study 2: Consumption in private settings

In the second study, I investigate whether people care about the moral values of a product's customer pool in anonymous settings. While consumption can not signal desirable characteristics to others, it might serve as self-signal.

According to the model, self-signaling motives can influence consumption choices in such settings if two conditions are satisfied. First, subjects must be insecure about their own moral values. Second, subjects must believe that a substantial share of a product's customers pool consists of people with certain characteristics in order for consumption to be informative about types. Observing the choice of only one consumer with a desirable or undesirable type, as done in Study 1, might not alter the perceived composition of a product's customers pool in they eves of consumers.[78] To accommodate both features, Study 2 asks for a substantially different design than Study 1. In Study 2, a product is adopted by many members of an undesirable group, likely altering the consumers' perceptions of the composition of the product's customer pool. To induce strong signaling motives, I opt for a group that is perceived as very undesirable, neo-Nazis. Conforming to the choices of neo-Nazis might constitute a bad signal about a subject's moral values, in particular racism. To increase subjects' uncertainty about their own racism, I threaten subjects' identities in the context of race before they make the product choices.

3.5.1 Experimental design

In the experiment, subjects choose between two products. Before they make their choices, I inform them about the choices made by others. I will refer to these other consumers as the *targets*, following the notation of Study 1. I randomly vary whether the targets are perceived to have neutral moral values (*neutral treatment*) or undesirable moral values (*undesirable treatment*). If subjects are concerned with their self-image, they might be relatively more attracted towards a product if its customer pool is perceived to consist of

[78] Remember that Study 1 was designed to create a strong correlation between types and choices in the eyes of the *observers*, not in the eyes of the consumers.

neutral types. Subjects make their choices in a double-blind setting to eliminate possible image concerns towards the experimenter.

3.5.1.1 The targets: a sample of neo-Nazis

To implement this study, I collected novel consumption data from neo-Nazis. The neo-Nazis play the role of targets in my experiment. I recruited 10 neo-Nazis on far-right extremist internet forums to participate in a short online survey. I convinced neo-Nazis to participate by paying high participation fees (20 Euros for 5 minutes) and offering an anonymous shipping option.[79] The survey consisted of eight binary product choices. One of the eight choice situations was randomly drawn, and participants received the product they chose in this choice situation. Subjects also filled out a short demographics questionnaire asking for gender, age, nationality, education and income.

Figure 14. Presentation of product choices to neo-Nazis

Notes: The colors show the symbols that might make a product more (red) or less (blue) attractive for neo-Nazis. "Bloch" is a common Jewish family name, "Milch," German for milk, is a symbol of the alt-right, and "Prügeli" is translated as "small baton," which describes the shape of the chocolate bars, but can also be understood as aggressive language.

For the purpose of my study, it is essential that one product is adopted by most neo-Nazis. To achieve this requirement, I linked some products with hidden symbols liked or disliked by neo-Nazis to increase or decrease, respectively, the attractiveness of these

[79] Hermes, a German postal service company, allows to send a packet to a pick-up point instead of a specific address. I also gave participants the option to send their payment to an address of their choice. I promised participants that all their information will be kept confidential, and that addresses will be deleted directly after their payment is shipped.

products. Some products had a 88 in their product numbers (a neo-Nazi symbol for HH, "Heil Hitler"), while others had the word "milk" (a recent symbol of the alt-right) or "rainbow" (a symbol for cultural diversity) in their names. One product, Camille Bloch Torino, is produced by a firm with Jewish background. I succeeded in that 9 out of 10 neo-Nazis preferred "Munz Praliné-Prügeli" milk chocolate over "Camille Bloch Torino" chocolate. Figure 14 shows how this choice situation was presented to neo-Nazis, and highlights all symbols that possibly make Munz more attractive than Camille Bloch. Table H2 in the Appendix gives all other choice situations and the corresponding neo-Nazi choices.

3.5.1.2 The products: two kinds of chocolates

In the laboratory, I investigate how the product adaptations of the neo-Nazis affect participants consumption choices. For products, I use the Munz chocolate and the Camille Bloch chocolate (see Figure 14). Using these products has multiple advantages. First, as discussed in the last paragraph, the Munz chocolate was very popular among neo-Nazis.

Second, there is no pre-existing association between the chocolate products and neo-Nazis, or any other specific political view. To provide evidence for this, I implemented a short online survey with 22 participants, drawn from the same subject pool that I use for my experiment. In the online survey, participants had to rate how useful and how popular different products, including the two chocolates, are for different groups of people, including neo-Nazis and students. I also asked participants whether they associate the consumers of these goods with a particular political position (left-wing extremist, left-wing, center, right-wing, right-wing extremist, no relation to political position). Participants neither associate Munz (or, Camille Bloch) with right-extremism, nor do they think that Munz is more useful for or more popular among neo-Nazis than Camille Bloch (see Figure F6 in the Appendix for details). Any connection between products and types therefore emerges through the type-composition of products' customer pools created in the laboratory.

A final advantage is that chocolate can be consumed in private. It is therefore unlikely that subjects avoid a product due to social-image concerns.

3.5.1.3 Laboratory experiment

In the laboratory, I first threaten subjects' identities. Next, subjects learn the choices of the targets. I manipulate the perception of the type-composition of products' customer pools by revealing the targets' moral values to some subjects, which is the key aspect of the

second experiment. Then, subjects make their product choice, fill out a survey and get paid in private.

Identity threat: I threaten subjects personal identities in the context of race to increase subjects' uncertainty about their own types. To do so, subjects complete an implicit association test (IAT)—a popular test in psychology to measure implicit racism—before they make their decision. I use the skin-tone IAT (Nosek, et al, 2007). Subjects learn their result in the IAT. The IAT very often reveals racism (65-82% of the population are implicit racists, according to Project Implicit[80]), and thereby threatens subjects' identities.[81] To avoid any social-signaling motives, IAT scores are not saved, and, as a result, neither the experimenter nor any other person sees the scores. To strengthen the identity threat, subjects receive detailed information about the interpretation of the results from the IAT at the beginning of the experiment.

Treatments: Next, subjects observe the choices of the 10 targets. Subjects are randomly assigned either to the undesirable or to the neutral treatment. The treatment is randomized within session. In both treatments, participants observe the choices of the neo-Nazis. However, the political ideology of targets is only revealed to the subjects in the undesirable treatment. In the neutral treatment, participants are only told that the targets were recruited on the internet. To keep the perception of the targets somewhat similar between treatments (in domains unrelated to moral values), participants in both treatments receive demographic information about the targets (distribution of gender, age and education). Figure H3 in the Appendix illustrates how this information is shown to subjects in both treatments.

Consumers' choices: Next, subjects choose between the two kinds of chocolate. As in Study 1, participants make 13 decisions between bundles of products and money (see Table 9). Unlike the first study, however, the choices are elicited in two stages. First, subjects make a binary choice between the two products. In this choice situation, none of the products is bundled with money. Then, they make the 13 decisions, where the choice between (Munz, CHF 0) and (Camille Bloch, CHF 0) is set to subjects' first-stage choices (but can be changed). At the end of the experiment, one of the 13 second-stage choices is randomly chosen to be implemented. (Unlike Study 1, there is only one round of consumption choice.)

Anonymity: To guarantee participants' anonymity, one participant is randomly selected to be the "monitor" at the beginning of each session. The monitor pays out

[80] https://implicit.harvard.edu/

[81] 45.23% of subjects reported in the survey at the end of the experiment that the IAT at least somewhat threatened their identity.

participants at the end of the study and does not participate in the experiment. The remaining participants each receive a random ID number hidden in an envelope. The experimenter cannot match the ID number to the participant. Each subject opens the envelope in private and enters this ID number in the computer terminal at which they sit. At the end of the study, each participant's monetary payoff and chosen product is placed in an envelope labeled only with the anonymous ID number. The monitor, who does not know the content in any of the envelopes, distributes the envelopes to the participants based one their ID numbers at the end of the study.

Survey: Before participants receive their payment, they fill out a survey. Given that it takes 20 to 30 minutes to fill the payment envelops after subjects made their product choices, the survey is very long to keep subjects occupied. Most importantly, I elicit subjects perception of the products, and test for experimenter demand effects.[82] I will give additional details on the survey in Section 6 when I discuss different explanations for my findings.

Sequence: First, subjects receive instructions explaining the entire experiment. While the instructions explain that they have to choose between two products ("product A" and "product B") in the 13 cases, it is not revealed what products A and B are. Instructions announce that participants will observe the choices of "10 participants from a previous study," but do not give any information about the types or choices of the targets. To strengthen the identity threat, instructions contain detailed information about the IAT. After reading the instructions, subjects answer a set of comprehension questions. Then, depending on the treatment they are told that the 10 other participants are either neo-Nazis or participants recruited on the internet. Next, it is revealed that the products correspond to Camille Bloch chocolate and Munz chocolate. At the same time, subjects learn that 9 out of 10 targets chose the Munz chocolate. This procedure avoids that subjects evaluate the products before they learn the preferences of targets. Next, subjects choose between the two products. Finally, participants fill out a survey and then receive their payment from the monitor.

Procedure: I conducted eleven sessions, each consisting of between 19 and 23 participants, resulting in a total of 243 participants (113 in the neutral treatment, 119 in the immoral treatment and 11 monitors). All sessions took place at the Laboratory for

[82] I added many questions to keep subjects occupied and to potentially inform future research. The questionnaire contains question about the popularity of the products among different groups of people, beliefs about the behavior of the other participants, political attitudes, expected use of the chocolate, familiarity with the chocolates, willingness to pay for a different chocolate, familiarity with the IAT and whether the IAT threatened subjects identities, image concerns, how prone subjects are to disgust, attitudes towards products popular among neo-Nazis and demographics. The complete questionnaire is in the Online Appendix.

experimental and behavioral economics at the University of Zurich, in December 2018. Participants were recruited using hroot (Bock, Baetge and Nicklisch, 2014) from the joint subject pool of the University of Zurich and the ETH. The experiment was implemented using z-Tree (Fischbacher, 2007) and, for the IAT, Minno.js (Zlotnick, Dzikiewicz and Bar-Anan, 2015). The Online Appendix supplies the instructions for the study.

3.5.2 Results

Do participants care about the moral values of the products' customers when they make consumption choices in this private setting? As in the analysis of Study 1, I will first look at the probability that subjects choose the Munz chocolate (the product popular among the targets) if none of the products are bundled with money.[83] Table 12, column (1) gives the estimates of a linear regression on the probability of choosing Munz over Camille Bloch when none of the chocolates are bundled with money on a treatment dummy (1 = undesirable treatment). In the neutral treatment, 56.4% of subjects chose Munz. In the undesirable treatment, however, only 43.1% of subjects chose Munz, a significant lower share (p=0.047). Table 12, column (2) demonstrates that this finding is robust to adding session fixed effects.

Table 12. Treatment effects

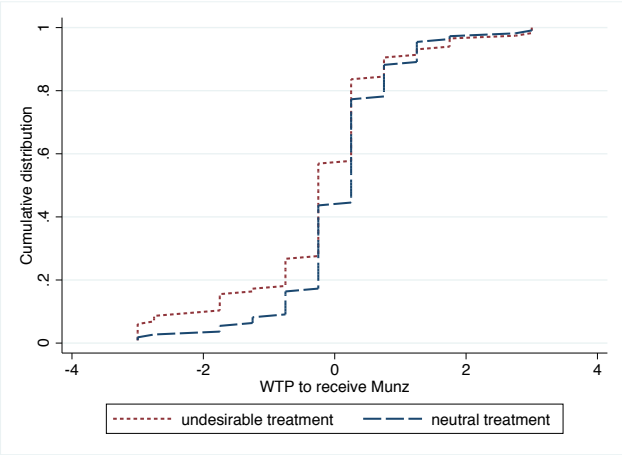
Dependent variable:	**Pr(Munz)**		**WTP Munz**	
	(1)	(2)	(3)	(4)
1 = undesirable treatment	-0.133** (-2.00)	-0.136** (-2.11)	-0.325** (-2.08)	-0.335** (-2.26)
Constant	0.564*** (11.87)		0.078 (0.69)	
Log(sigma)			1.169*** (20.12)	1.108*** (20.14)
N	226	226	226	226
Session Fixed Effects	No	Yes	No	Yes

*Notes: (1) and (2): Linear regressions of probability to choose Munz chocolate when none of the two products are connected with any payment on a treatment dummy. Robust standard errors are used. (3) and (4): Tobit regressions (left-censored at CHF -3, n=5; right-censored at CHF +3, n=9) of willingness to pay to receive the Munz chocolate instead of the Camille Bloch chocolate on a treatment dummy. t-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

[83] Six participants made product choices that are non-monotone in money. These individuals are excluded from the analysis. Table F3 in the Appendix shows that results do not change if these observations are kept.

The data allows me to calculate subjects' willingness to pay to receive the Munz chocolate instead of the Camille Bloch chocolate.[84] Note that this variable is left-censored at CHF -3 (5 observations) and right-censored at CHF 3 (9 observations). Figure 15 shows the cumulative distribution of the willingness to pay for both treatments. Most participants have a willingness to pay in between CHF -1 and CHF 1. The figure indicates that the treatment shifted the willingness to pay distribution.

Figure 15. Distribution willingness to pay, Study 2



Notes: Cumulative distributions of subjects willingness to pay to receive Munz product for both treatments.

Table 12, columns (3) gives the estimates of a Tobit regression on the willingness to pay on a treatment dummy. Subjects in the undesirable treatment are willing to pay CHF 0.325 less for the product popular among the targets than subjects in the neutral treatment (p=0.038). Column (4) demonstrates that this finding is robust to adding session fixed effects. Note that the effect size is about 10 percent of the price of the chocolates.[85] This is a substantial effect given that European net margins for food processing and food retail are 6.76% and 1.67%, respectively (Damodaran, 2019).

[84] My data is discrete due to use of the list method. For a subject, who choses Camille Bloch in the choice situation "(Camille Bloch, CHF 1.50) or (Munz, CHF 0)", but choses Munz in the choice situation "(Camille Bloch, CHF 1) or (Munz, CHF 0)", the difference in monetary value must be in [CHF 1, CHF 1.5]. I set this difference to CHF 1.25.

[85] When I hired subjects, I ask students to not sign up for this study in case they are allergic to chocolate. I did so because a chocolate choice of a participant that can not eat chocolate does not reveal information about the participant's type to herself. Nevertheless, 15 students indicated in the survey that they neither can eat the Munz nor the Camille Bloch chocolate. If I exclude these subjects from the analysis, treatment effects increase; treatment effects in Table 12 are then -0.166 (t=-2.43, p=0.016) for (1), 0.160 (t=-2.38, p=0.018) for (2), -0.415 (t=-2.60, p=0.010) for (3) and -0.413 (t=-2.72, p=0.007) for (4).

In this second study, I find that the type-composition of customer pools also matter for private consumption. In the following, I will discuss different motives that potentially could explain the behaviors documented in the two studies.

3.6 Alternative explanations

In both studies, I find that consumers care about the moral values of others that chose a product, as predicated by models of identity signaling. In the following, I discuss whether an explanation other than identity signaling could produce my findings. Understanding the motives of consumer behavior is potentially important for policy implications of my work; identity signaling, for example, often results in suboptimal consumption behaviors (Frank, 1985; Ireland, 1994; Moav and Neeman, 2010, 2012).

Perception of the products: Changing the type-composition of a product's customer pool might change the perception of the product's qualities or the producer's qualities, for example through forms of social learning (Bikhchandani, Hirshleifer and Welch, 1998). Such treatment differences in the perception of products might then be reflected in the consumption choices.

To investigate whether there are treatment differences in such perceptions, I complement my studies with survey questions. I elicit participants' product perceptions on different quality dimensions (including price, processing quality, and quality of raw materials) and of the moral values of the producers. I do not find treatment differences in most dimensions of product quality (see Table F6 and Table F7 in the Appendix). Moreover, there are no treatment differences in the perception of the producers' moral values: in Study 1, subjects do not think that the producer of the product that is adopted by the target with undesirable moral values promotes conservative Christian values (Wilcoxon rank-sum, $p=0.611$; see Table F6). In Study 2, subjects in the undesirable treatment do not think that Munz, the product adopted by neo-Nazis, promotes right-wing extremism or discriminates against minorities (Wilcoxon rank-sum, $p=0.800$, $p=0.406$, respectively; see Table F7).

Disgust: Associating a product with neo-Nazis, as done in Study 2, might remind the buyer unpleasantly Nazis and their actions, resulting in feelings of disgust when consuming the product—in line with negative contagion (Rozin, Millman and Nemeroff, 1986; Nemeroff and Rozin, 1994)—thereby devaluating the product.

In the survey, I ask subjects whether they feel disgusted when they think about eating the product chosen by their target. There are no treatment differences in subjects' responses in both Study 1 and Study 2 (Wilcoxon rank-sum, p=0.967, p=0.422, respectively; see Table F6 and F7).[86]

Similarity to the target: People might be more willing to conform to behaviors of others that have similar personality traits, values or group affiliations than to others that are different in these dimensions. There is some evidence for such an effect for moral behavior (Dimant, 2019; Fatas, Hargreaves Heap and Rojo Arjona, 2018). Treatment differences in the similarity between consumers and targets could therefore potentially explain the treatment differences in consumer behavior.

My data allows me to measure the similarity between the moral values of consumers and targets. For Study 1, I look at the similarity in donations to Zukunft CH.[87] Consumers in the undesirable treatment are on average more similar to their targets than the consumers in the desirable treatment (t=2.99; p=0.003).[88] Therefore, if participants are more willing to conform to more similar targets, conformity should be more common in the undesirable treatment than in the desirable treatment. I find the opposite pattern. For Study 2, I use subjects' political position, racism and views on right-wing and left-wing extremism to measure similarity: subject that are right-wing and hold racist views are more similar to the neo-Nazis than subjects that are left-wing and opposed to racism. In both studies, the similarity measures can not account for the treatment differences; if anything, controlling for similarity increases treatment effects (see Tables F4 and F5 in the Appendix).

Experimenter Demand Effects: One may be worried about experimenter demand effects, in particular given my use of neo-Nazis as undesirable types in Study 2. However, experimenter demand effects can hardly explain why there is a treatment effect for moral values, but not for intelligence; if experimenter demand effects were present, they likely would occur in both the moral values round and the intelligence round. In Study 2, I attempt to test for experimenter demand effects, building on the method developed by de Quidt, Haushofer and Roth (2018). In the survey, subjects are asked to submit an offer (in [0 CHF,

[86] In the survey, I also ask subjects how they plan to use the chocolates in case they receive them: whether they eat it, whether they give it to someone else, and whether they throw the chocolate away. If disgust plays an important role, subjects might be less likely to eat the Munz chocolate, but more likely to give it away or throw it away. I do not find a treatment difference in any of the three variables (linear probability model; t=0.25 (p=0.801), t=1.54 (p=0.126), t=-0.94 (p=0.346), respectively).

[87] Note that in the moral values round, targets in the desirable and undesirable treatment only differ in their donation to Zukunft CH.

[88] I calculated the similarity as $1 - |\text{donation} - \text{target's donation}|/6$ and regressed this variable on a treatment dummy. The coefficient estimate is 0.16.

6.05 CHF]) to buy an USB stick, incentivized by the BDM. Half of the subjects are assigned to a demand treatment. In the demand treatment, I add the sentence "We expect that participants who are shown these instructions will specify a lower maximum price than they normally would." As de Quidt, Haushofer and Roth (2018) argue, such a sentence induces experimenter demand effects and allows to test whether subjects are prone to experimenter demand effects in a particular setting. I do not find a treatment differences in the willingness to pay for the USB stick (Tobit regression, coeff.=-0.034, t=-0.11, p=0.916). Given that this setting is similar to the consumption decisions in the main part of my experiment, it seems unlikely that experimenter demand effects drive my results.

Unlike all other explanation, identity signaling can explain all aspects of behaviors observed in the experiments.

3.7 Conclusion and discussion

This paper studies whether consumers care about the characteristics of others that buy a product when the make consumption choices. In laboratory experiments, I vary, by treatment, whether the customer pool of a product consists of people with desirable or undesirable characteristics. I find that participants are willing to pay more for a product if it is popular among people with desirable *moral values* than if it is popular among people with undesirable moral values. I show this in a setting where choices are observed by others, and in a double blind setting.

In addition to moral values, I investigate whether consumers care about the intelligence of others that buy a product. In contrast to moral values, I do not find evidence for such an effect. These findings suggest that, at least in in the contexts studied in this paper, subjects care more about the public perception of their moral values than about the public perception of their intelligence.

These findings can explain why many firms want customers with undesirable moral values—for example, alt-rights, neo-Nazis and Hooligans—*not* to buy their products. Adaptations by such customers can decrease other customers' willingness to pay for the firms' products. As a result, it can be rational for firms to forgo profits generated through customers with undesirable moral values in order to make more profit with other customers.

One contribution of this paper is to deliver important evidence in support of major models of identity signaling and consumption. These models predict that consumers want to

signal that they have desirable characteristics (or, "types") by avoiding products popular among people with undesirable characteristics and by conforming to the product choices of people with desirable characteristics. I present evidence that consumers indeed care about the types of others that consume a product as postulated in these models. From a theoretical point of view, my work also provides evidence for a violation of a fundamental assumption in consumer theory, that is, that consumers do not care about the consumption bundles of others. Contrary to this assumption, subjects in my studies avoid consumption choices popular among people with undesirable moral values.

My work also has potentially important policy implications. For instance, some social groups have adopted suboptimal consumption behaviors, such as unhealthy products (Guendelman, Cheryan, and Monin, 2011), food with low caloric intake consumption (Atkin, 2016) or products that harm the environment (Minton, Johnson and Liu, 2019). I present evidence that consumers are attracted towards products that are popular among people with desirable characteristics. Policies that target members of groups that are seen as having very desirable (or, very undesirable) characteristics might therefore be successful in changing suboptimal consumption pattern of the entire group.

While I focus my investigation on consumer behavior, understanding how people's choices depend on the choice of other types is also important for other applications, for example, for the literature on acting white.[89] This literature suggests that studying hard is seen as a "white action" by black students, and avoided to signal group-attachment (e.g., Fordham and Ogbu, 1986). Existing studies on acting white are limited to correlations in field-data and are the subject of controversial discussion (e.g., Austen-Smith and Fryer, 2005).[90] My investigation provides a empirical framework to study how people's choices are affected by the choices of types that vary along several dimensions and enhances our general understanding of the psychological forces behind conformity and disconformity in such contexts.

[89] There is also related evidence that women avoid educational choices that are associated with men (Cheryan et al., 2009) and that low-income groups and ethnic minorities in the U.S. associate health concerns with white, middle-class Americans and adopt unhealthy behaviors (Oyserman, Fryberg and Yoder, 2007).

[90] Bursztyn, Egorov and Jensen (2019) provide causal evidence on how social-image concerns affect educational activities in different school environments. They do not study how people's choices depend on the choice of other types of people.

References

Andreoni, J. and Bernheim, B. D. (2009) "Social Image and the 50–50 Norm: a Theoretical and Experimental Analysis of Audience Effects," Econometrica, 77(5): 1607–1636.

Akerlof, G. and Kranton, R. E. (2000) "Economics and Identity," Quarterly Journal of Economics 115(3), 715-753.

Atkin, D. (2016) "The Caloric Costs of Culture: Evidence from Indian Migrants," American Economic Review, 106(4): 1144-1181.

Atkin, D., Colson-Sihra, E. and Shayo, M. (2019) "How Do We Choose Our Identity? A Revealed Preference Approach Using Food Consumption," NBER Working Paper No. 25693.

Austen-Smith, D. and Fryer, R. G. (2005) "An Economic Analysis of "Acting White"," Quarterly Journal of Economics, 120(2): 551-583.

Bagwell, L. S. and Bernheim, B. D. (1996) "Veblen Effects in a Theory of Conspicuous Consumption," American Economic Review, 86(3): 349-373.

BBC News (2005) "Burberry versus The Chavs," October 28.

Bénabou, R. and Tirole, J. (2011). "Identity, Morals, and Taboos: Beliefs as Assets," Quarterly Journal of Economics, 126: 805-855.

Bernheim, B. D. (1994) "A Theory of Conformity," Journal of Political Economy, 102(5): 841-877.

Bernheim, B. D. and Exley, C. L. (2015) "Understanding Conformity: An Experimental Investigation," working paper.

Bertrand, M. and Kamenica, E. (2018) "Coming apart? Cultural distances in the United States over time," working paper.

Berger, J. and Heath, C. (2007) "Where Consumers Diverge from Others: Identity Signaling and Product Domains," Journal of Consumer Research, 34: 121–134.

Berger, J. and Heath, C. (2008) "Who Drives Divergence? Identity Signaling, Outgroup Dissimilarity, and the Abandonment of Cultural Tastes," Journal of Personality and Social Psychology, 95(3): 593-607.

Berger, J. and Rand, L. (2008) "Shifting Signals to Help Health: Using Identity Signaling to Reduce Risky Health Behaviors," Journal of Consumer Research, 35(3): 509-518.

Bessendorf, A. and Gans, S. (2015) "From Cradle to Cane: The Cost of Being a Female Consumer. A Study of Gender Pricing in New York City," New York City Department of Consumer Affairs.

Bigenho, J. and Martinez, S. (2018) "Social Comparisons in Peer Effects," working paper.

Bikhchandani, S., Hirshleifer, D. and Welch, I. (1998) "Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades," Journal of Economic Perspectives, 12(3): 151-170.

Bloch, F., Rao, V. and Desai, S. (2004) "Wedding celebrations as conspicuous consumption: signaling social status in rural India," Journal of Human Resources, 39: 675-695.

Bloomberg (2019) "Patagonia Is Cracking Down on the Wall Street Uniform," April 3.

Bock, O., Baetge, I., and Nicklisch, A. (2014) "hroot: Hamburg registration and organization online tool," European Economic Review, 71: 117-120.

Bourdieu, Pierre (1984) "Distinction: A Social Critique of the Stratification of Taste," Cambridge, MA: Harvard University Press.

Bursztyn, L., Egorov, G. and Fiorin, S. (2017) "From Extreme to Mainstream: How Social Norms Unravel," working paper.

Bursztyn, L., Egorov, G. and Jensen, R. (2019) "Cool to be Smart or Smart to be Cool? Understanding Peer Pressure in Education," Review of Economic Studies, 86(4): 1487-1526.

Bursztyn, L., Ferman, B., Fiorin, S., Kanz, M. and Rao, G. (2018) "Status Goods: Experimental Evidence from Platinum Credit Cards," Quarterly Journal of Economics, 133(3): 1561-1595.

Business Insider (2013) "Abercrombie & Fitch Refuses To Make Clothes For Large Women," May 3.

Carbajal, J. C., Hall, J. and Li, H. (2016) "Inconspicuous conspicuous consumption," working paper.

Charles, K.K., Hurst, E. and Roussanov, N. (2009) "Conspicuous consumption and race," Quarterly Journal of Economics, 124: 425-467.

Cheryan, S., Plaut, V. C., Davies, P. G. and Steele, C. M. (2009) "Ambient Belonging: How Stereotypical Cues Impact Gender Participation in Computer Science," Journal of Personality and Social Psychology, 97(6): 1045-1060.

Clingingsmith, D. and Sheremeta, R. M. (2018) "Status and the demand for visible goods: experimental evidence on conspicuous consumption," Experimental Economics, 21: 877-904.

Corneo, G. and Jeanne, O. (1997) "Conspicuous consumption, snobbism and conformism," Journal of Public Economics, 66: 55-71.

Cosaerts, S. (2018) “Revealed Preferences for Diamond Goods,” American Economic Journal: Microeconomics, 10(2): 83-117.

Damodaran, A. (2019) “Profit margins (net, operating and EBITDA), Western Europe.” http://people.stern.nyu.edu/adamodar/New_Home_Page/datacurrent.html

Delgado, M. S., Harriger, J. L. and Khanna, N. (2015) “The value of environmental status signaling,” Ecological Economics, 111: 1-11.

de Quidt, J., Haushofer, J. and Roth, C. (2018) “Measuring and Bounding Experimenter Demand,” American Economic Review, 108(11): 3266-3302.

Dimant, E. (2019) “Contagion of pro- and anti-social behavior among peers and the role of social proximity,” Journal of Economic Psychology, 73: 66-88.

Dubé, J., Luo, X., Fang Z. (2017) “Self-Signaling and Prosocial Behavior: A Cause Marketing Experiment,” Marketing Science, 36(2): 161-186.

Dunn, L., White, K. and Dahl, D. W. (2012) “That Is So Not Me: Dissociating from Undesired Consumer Identities,” in Ruvio, A. A. and Belk, R. W. eds., The Routledge Companion to Identity and Consumption, New York: Routledge.

Economist (2006) “Bubbles and bling,” May 8.

Ellis, L., and Ficek, C. (2001) “Color preferences according to gender and sexual orientation,” Personality and Individual Differences, 31: 1375-1379.

Fatas, E., Hargreaves Heap, S. P., and Rojo Arjona, D. (2018) “Preference conformism: An experiment,” European Economic Review, 105: 71-82.

Fischbacher, U. (2007) “z-Tree: Zurich toolbox for ready-made economic experiments,” Experimental Economics, 10(2): 171–178.

Fordham, S. and Ogbu, J. U. (1986) ‘‘Black Students’ School Success: Coping with the ‘Burden of Acting White’,’’ Urban Review, 18: 176–206.

Frank, R. H. (1985) “The demand for unobservable and other positional goods,” American Economic Review, 75(1): 101-116.

Friedrichsen, J. (2018) “Signals sell: Product lines when consumers differ both in taste for quality and image concern,” working paper.

Friedrichsen, J., Engelmann, D. (2018) “Who cares about social image?,” European Economic Review, 110: 61-77.

Gao, L. S., Wheeler, C. and Shiv, B. (2009) “The ‘Shaken Self’: Product Choices as a Means of Restoring Self-View Confidence,” Journal of Consumer Research, 36(1): 29-38.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L. (2017) "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States," PNAS, 114(50): 13108–13113.

Gioia, F. (2017) "Peer effects on risk behaviour: the importance of group identity," Experimental Economics, 20(1): 100-129.

Glamour (2017) "These are the fashion designers that will flat out NOT dress Melania Trump," April 13.

Glazer, A. and Konrad, K. A. (1996) "A Signaling Explanation for Charity," American Economic Review, 86(4): 1019-1028.

Gneezy, A., Gneezy, U., Riener, G., Nelson, L. D. (2012) "Pay-what-you-want, identity, and self-signaling in markets," PNAS, 109(19): 7236-7240.

Griskevicius, V., Tybur, J. M., and Van den Bergh, B. (2010) "Going Green to Be Seen: Status, Reputation, and Conspicuous Conservation," Journal of Personality and Social Psychology, 2010, Vol. 98, No. 3, 392–404.

Guendelman, M. D., Cheryan, S. and Monin, B. (2011) "Fitting In but Getting Fat: Identity Threat and Dietary Choices Among U.S. Immigrant Groups," Psychological Science, 22(7): 959-967.

Hänni, S. and Lichand, G. (2019) "Harming to Signal," working paper.

Handelsblatt (2014) "Wie 'LoNSDAle' sein Nazi-Problem löst," March 26.

Heffetz, O. (2011) "A test of conspicuous consumption: visibility and income elasticities," Review of Economics and Statistics, 93(4): 1101-1117.

Heffetz, O. (2018) "Expenditure Visibility and Consumer Behavior: New Evidence," NBER working paper 25161.

Hopkins, E. and Kornienko, T. (2004) "Running to Keep in the Same Place: Consumer Choice as a Game of Status," American Economic Review, 94(4): 1085-1107.

Independent (2017) "Depeche Mode call Richard Spencer a 'c***' after white supremacist branded them 'official band of the alt-right'," March 15.

Ireland, N. J. (1994) "On Limiting the Market for Status Signals," Journal of Public Economics, 53: 91-110.

Ireland, N. J. (1998) "Status-seeking, income taxation and efficiency," Journal of Public Economics, 70: 99-113.

Kapner, S. and Chinni, D. (2019) "Are Your Jeans Red or Blue? Shopping America's Partisan Divide," Wall Street Journal.

Karni, E. and Schmeidler, D. (1990) "Fixed Preferences and Changing Tastes," American Economic Review, Papers and Proceedings, 80(2): 262-267.

Kaus, W. (2013) "Conspicuous consumption and 'race': Evidence from South Africa," Journal of Development Economics, 100: 63–73.

Khamis, M., Prakash, N. and Siddique, Z. (2012) "Consumption and social identity: Evidence from India," Journal of Economic Behavior & Organization, 83: 353-371.

Khan, R., Misra, K. and Singh, V. (2013) "Ideology and Brand Consumption," Psychological Science, 24(3): 326-333.

Kidwell, B., Farmer, A. and Hardesty, D. M. (2013) "Getting Liberals and Conservatives to Go Green: Political Ideology and Congruent Appeals," Journal of Consumer Research, 40(8): 350-567.

Kim, S. and Rucker, D. D. (2012) "Bracing for the Psychological Storm: Proactive versus Reactive Compensatory Consumption," Journal of Consumer Research, 39(4): 815-830.

Krupka, E. and Weber, R. A. (2009) "The focusing and informational effects of norms on pro-social behavior," Journal of Economic Psychology, 30(3): 307-320.

Kuksov, D. (2007) "Brand Value in Social Interaction," Management Science, 53(10): 1634-1644.

Kuksov, D., Shachar, R. and Wang, K. (2013) "Advertising and Consumers' Communications," Marketing Science, 32(2): 294-309.

Kuksov, D. and Wang, K. (2013) "A Model of the 'It' Products in Fashion," Marketing Science, 32(1): 51-69.

Lahno, A. and Serra-Garcia, M. (2015) "Peer effects in risk taking: Envy or conformity?," Journal of Risk and Uncertainty, 50(1): 73-95.

Leibenstein, H. (1950) "Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand," Quarterly Journal of Economics, 64: 183-207.

McConnell, C., Margalit, Y., Malhotra, N. and Levendusky, M. (2018) "The Economic Consequences of Partisanship in a Polarized Era," American Journal of Political Science, 62(1): 5-18.

McManus, T. C. and Rao J. M. (2015) "Signaling smarts? Revealed preferences for self and social perceptions of intelligence," Journal of Economic Behavior & Organization, 110, 106-118.

Moav, O. and Neeman, Z. (2010) "Status and Poverty," Journal of the European Economic Association, 8(2–3): 413-420.

Moav, O. and Neeman, Z. (2012) “Saving Rates and Poverty: the Role of Conspicuous Consumption and Human Capital,” Economic Journal, 122: 933-956.

Mochon, D., Norton, M. I. and Ariely, D. (2012) “Bolstering and restoring feelings of competence via the IKEA effect,” International Journal of Research in Marketing, 29: 363–369.

Minton, E. A., Johnson, K. A., and Liu, R. L. (2019) “Religiosity and special food consumption: The explanatory effects of moral priorities,” Journal of Business Research, 95:442-454.

Nemeroff, C. and Rozin, P. (1994) “The Contagion Concept in Adult Thinking in the United States: Transmission of Germs and of Interpersonal Influence,” Ethos, 22(2): 158-186.

New York Times (2018) “Cambridge Analytica Used Fashion Tastes to Identify Right‐Wing Voters,” November 29.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ratliff, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G. and Banaji, M. R. (2007) “Pervasiveness and correlates of implicit attitudes and stereotypes,” European Review of Social Psychology, 18: 36-88.

Oyserman, D., Fryberg, S. A. and Yoder, N. (2007) “Identity-Based Motivation and Health,” Journal of Personality and Social Psychology, 93(6): 1011-1027.

Pesendorfer, W. (1995) “Design Innovation and Fashion Cycles,” American Economic Review, 85(4): 771-792.

Rao, E. S. and Schaefer, R. (2013) “Conspicuous Consumption and Dynamic Pricing,” Marketing Science, 32(5): 786-804.

Roos, J. and Shachar, R. (2013) “When Kerry Met Sally: Politics and Perceptions in the Demand for Movies,” Management Science, 60(7): 161716-31.

Rozin, P., Millman, L. and Nemeroff, C. (1986) “Operation of the Laws of Sympathetic Magic in Disgust and Other Domains,” Journal of Personality and Social Psychology 50(4): 703-712.

Rucker, D. D. and Galinsky, A. D. (2008) “Desire to Acquire: Powerlessness and Compensatory Consumption,” Journal of Consumer Research, 36: 257-610.

Ruvio, A. A. and Belk, R. W. (2013) “The Routledge companion to identity and consumption,” London: Routledge.

Sexton, S. E. and Sexton, A. L. (2014) "Conspicuous conservation: The Prius halo and willingness to pay for environmental bona fides," Journal of Environmental Economics and Management, 67(3): 303-317.

Simmel, G. (1957) "Fashion," American Journal of Sociology, 62(6): 541-558. (Reprinted from International Quarterly, 1904, 10: 130-155.)

Trigg, A. B. (2001) "Veblen, Bourdieu, and Conspicuous Consumption," Journal of Economic Issues, 35(1): 99-115.

Veblen, T. (1994) "The theory of the leisure class: An economic study of institutions," New York: Dover Publications. (Reprinted from London: Unwin Books, 1899.)

Vikander, N. (2017) "Advertising to Status-Conscious Consumers," working paper.

Washington Post (2017) "What happens when neo-Nazis hijack your brand," November 16.

Wernerfelt, B. (1990) "Advertising Content When Brand Choice is a Signal," Journal of Business, 63(1): 91-98.

White, K. and Dahl, D. W. (2006) "To Be or Not Be? The Influence of Dissociative Reference Groups on Consumer Preferences," Journal of Consumer Psychology, 16(4): 404-414.

White, K. and Dahl, D. W. (2007) "Are All Out-Groups Created Equal? Consumer Identity and Dissociative Influence," Journal of Consumer Research, 34(4): 525-536.

Zafar, B. (2011) "An experimental investigation of why individuals conform," European Economic Review, 55(6): 774-798.

Zimmermann, F. (forthcoming) "The Dynamics of Motivated Beliefs," American Economic Review.

Zlotnick, E., Dzikiewicz, A. J. and Bar-Anan, Y. (2015). Minno.js, computer software.

Chapter 4: On self-serving strategic beliefs

Joint with Nadja Ging-Jehli and Roberto A. Weber

Abstract

We experimentally study whether individuals adopt negative beliefs about others' intentions to justify egoistic behavior. In contrast with Di Tella, et al. (2015), our first study finds no evidence that individuals engage in "strategic cynicism." We reconcile the discrepancy, using Di Tella, et al.'s, data, a simple model of belief manipulation and a novel experiment that replicates and extends Di Tella, et al. Across three datasets, we find no evidence of negatively biased beliefs. However, Di Tella, et al.'s, results and our data indicate that those with a greater incentive to view others' intentions cynically exhibit relatively less positive beliefs.

Citation

A version of this paper is published as: Ging-Jehli, N. R., Schneider, F. H. and Weber, R. A. (2020) "On self-serving strategic beliefs," Games and Economic Behavior, 2020, Vol. 122, pp. 341-353.

4.1 Introduction

Considerable evidence indicates that decision makers confronted with tradeoffs between egoistic and social considerations, such as fairness and equality, will rely on justifications to prioritize the former while avoiding the impression that they are acting selfishly (Dana, Weber and Kuang, 2007; Hamman, Loewenstein and Weber, 2010; Gino, Norton and Weber, 2016; Grossman and van der Weele, 2017). This includes engaging in self-serving belief manipulation, whereby actions that are personally beneficial can be justified by changing one's beliefs or perceptions of what is fair (Babcock and Loewenstein, 1997; Konow, 2000), a product's quality (Chen & Gesche, 2017; Gneezy, Saccardo, Serra-Garcia, and van Veldhuizen, 2018) or the likely outcomes of a random process (Haisley and Weber, 2010; Exley, 2016).[91]

One important, but largely unexplored, context for self-serving belief manipulation is in strategic settings where individuals form beliefs about an opponent's likely behavior. It has long been recognized that beliefs about other players' actions and intentions can play a central role in prosocial behavior, with a positive relationship between the belief that others will act unkindly and one's own egoistic behavior (e.g., Rabin, 1993; Levine, 1998; Fischbacher, Gächter and Fehr, 2001). Strategic beliefs are typically assumed to be determined by the structure of the game and beliefs about others' preferences or rationality. However, in light of the apparent ease with which people bias their beliefs in self-serving ways in other contexts, it seems plausible that they may similarly bias their beliefs about others' actions when doing so can justify acting in a selfish way that harms others.[92] Indeed, a recent paper by Di Tella, Perez-Truglia, Babino and Sigman (2015) provides evidence consistent with the idea that people engage in such "strategic cynicism." Specifically, they

[91] Such self-serving interpretations are related to the concept of "motivated reasoning" from psychology (Kunda, 1990). Models introducing self-deception and self-image concerns to economics include Akerlof and Dickens (1982), Rabin (1994), Akerlof and Kranton (2000), Bénabou and Tirole (2006, 2011), Bénabou (2013). There is also evidence for self-serving belief manipulation about other desired qualities, like one's abilities (Möbius, Niederle, Niehaus and Rosenblat, 2017; Zimmerman, 2018), beauty (Eil and Rao, 2011), honesty (Mazar, Amir, and Ariely 2008; Shalvi, Gino, Barkan, and Ayal, 2015) and about desired future life events (Irwin, 1953; Mayraz, 2013).

[92] There are many contexts in which adopting cynical beliefs about others' likely actions may be strategically desirable for an individual constrained to act morally. For instance, an employer who can benefit by laying off a worker may find it easier to do so if she adopts the belief that the employee is likely committing acts that merit firing. A national leader intent on seizing land from a neighboring country may find this easier to justify under the belief that the other country intends to act aggressively. A US President may find it easier to justify firing a special counsel investigating him for misconduct if he convinces himself that the investigation is a WITCH HUNT! Trivers (2011) discusses an alternative motive to engage in self-deception in strategic situations: deceiving oneself about one's own qualities, such as ability, might be an effective strategy to deceive others (see also Schwardmann and van der Weele (2016), for related experimental evidence).

demonstrate that people with a greater opportunity to take from another person believe that this opponent is more likely to act in a greedy and harmful manner.

Our study investigates the phenomenon of strategic self-deception, although we initially approach this question in a different manner from Di Tella, et al. Rather than testing whether people with a *greater* incentive to take from others adopt *relatively* more negative beliefs about these opponents, as they do, our focus is on whether people with the opportunity to take from others adopt beliefs that are biased in comparison to two reasonable objective standards: the actual empirical behavioral frequency of opponents' behavior and the beliefs of neutral outsiders with no incentive to view others self-servingly. That is, we test the extent to which individuals with an incentive to engage in strategic cynicism adopt beliefs that are negatively biased in *absolute* terms. In contrast, Di Tella, et al., study a *relative* form of this bias, investigating whether one group's beliefs are more negative—or, critically, less positive—than those of another group.

Our results show that this distinction is important. We find evidence consistent with the relative bias documented by Di Tella, et al. However, we also show in two novel studies and in Di Tella, et al.'s, own data that there is no evidence of negatively biased strategic beliefs in absolute terms. In fact, across all three studies, individuals with an incentive to take from others—and, therefore, with an incentive to engage in strategic cynicism—actually hold highly accurate beliefs that are close to both the actual behavior of their counterparts and to the beliefs of neutral observers. The only bias, relative to these objective and neutral standards, in all three data sets lies in the beliefs of individuals in Di Tella, et al.'s, design with a *low* ability to take from their counterpart; these people exhibit overly *positive* beliefs about their counterpart's likely action. Thus, to the extent that absolute bias exists in people's beliefs about a counterpart's actions, it appears to be one of positivity rather than cynicism.

Our first study, which we conducted prior to knowing about Di Tella, et al.'s, related work, uses a game that we refer to as the "pre-emptive taking game." In this game, a pair of players—say, "Ann" and "Bob"—both start off with the same wealth endowment. Ann first decides how much to take from Bob. Bob then decides how much to take from Ann. A key feature is that Bob's ability to take from Ann increases in the amount of money he has remaining after Ann's taking decision. That is, by taking from Bob, Ann both increases her earnings and reduces his opportunity to act in a selfish and harmful manner. Thus, Ann's taking decision may naturally be influenced by whether or not she thinks Bob will use his remaining money to harm Ann. But, if Ann feels constrained to act "fairly," the game also

creates an incentive for Ann to manipulate her beliefs about Bob's likely action, since this gives her a justification for taking more under the guise of self-protection.

Our main purpose in this study is to test for a bias in the beliefs of subjects in the role of Ann. Therefore, we directly elicit such beliefs about the amount that Bob will take if given the opportunity. We compare this to beliefs elicited from neutral third parties who have no incentive to engage in strategic cynicism. Our hypothesis is that Ann's incentive to justify taking by adopting a cynical belief about Bob's likely behavior will lead her to self-servingly negatively bias these beliefs. Surprisingly, however, in light of other instances in which people seem to engage in self-deception in non-strategic contexts, we find no difference between the two sets of beliefs. The two beliefs are virtually identical and very close to the true empirical frequencies. This suggests, at the least, that there are limitations in people's ability to manipulate their beliefs about a strategic opponent.

This finding also contrasts with those of Di Tella, et al., who argue that their experimental evidence shows that individuals form biased beliefs and convince themselves that their counterparts are more likely act egoistically than they actually are.[93] Di Tella, et al., support this conclusion with an experiment using a game, labeled the "corruption game," that shares features with our pre-emptive taking game. In this game, Ann and Bob again start with an identical number of tokens and Ann similarly decides how many tokens to take from Bob. In the corruption game, however, Bob simultaneously makes a binary decision whether to act "corruptly," by taking a side payment that increases Bob's wealth while lowering the value of the tokens. Thus, Ann can justify taking more tokens from Bob as a fair action if she thinks that he will act corruptly. The experiment manipulates Ann's ability to take tokens from Bob and finds that subjects in the role of Ann adopt more pessimistic beliefs when they have the ability to take more tokens. Thus, individuals seem to respond to the incentive to take more from their counterpart by engaging in strategic cynicism.

At first, these two results seem to offer conflicting evidence. Our first study finds no evidence of strategic cynicism. In contrast, Di Tella, et al., conclude that people engage in self-serving belief manipulation. We reconcile this apparent inconsistency using the data from Di Tella, et al.'s, experiment, a simple model of strategic self-deception and an additional novel experiment. We show that the two sets of results are actually highly

[93] Specifically, in a context where the actual proportion of egoistic counterparts is p_0, Di Tella, et al., argue that a self-servingly biased decision maker "may form a biased belief [...] instead of correctly remembering a proportion of p_0 of low-type, the individual may try to convince herself that the proportion was actually $p > p_0$" (p. 3437).

consistent and, in doing so, provide important evidence on the nature of strategic self-deception.

First, we show that a closer inspection of Di Tella, et al.'s, data reveals that their results, like ours, do not actually show any absolute cynicism or bias on the part of individuals with an incentive to act egoistically. In fact, the subjects in their experiment with the greater incentive to engage in strategic self-deception provide belief estimates very close to the actual frequency of egoistic behavior by their counterparts. In contrast, those subjects with a low incentive to engage in strategic cynicism exhibit the most bias, but in the direction of believing that their opponents will be *less egoistic* than they actually are. Thus, to the extent that an empirical bias exists in Di Tella, et al.'s, data, it seems not to be one of cynicism, but rather one of optimism and positivity that arises only among those with little ability to take from their opponent.

We next show that, theoretically, this positivity bias can be explained by a simple model that serves as a stylized representation of the games in both ours and Di Tella, et al.'s, experiments. In the model, Ann derives utility from her own and Bob's payoffs and this utility is increasing in Bob's kindness. When Ann has the opportunity to take from Bob, she has an incentive to reduce her belief regarding Bob's kindness; this diminishes the loss in utility she experiences by taking from him. However, in this setting individuals also have an incentive to form another kind of motivated belief—to convince themselves that the other player is kind and deserves any payoff she receives. The net result of these two opposing tendencies is an absolute bias in the direction of positivity, which is consistent with a general tendency for distorted beliefs to lie in the direction of positivity and optimism, rather than the opposite (Bénabou and Tirole, 2016). This simple theoretical analysis provides a basis for two phenomena we observe in our first experiment and in Di Tella, et al.'s, data. First, in absolute terms, biases about others' actions will lie in the direction of positivity rather than cynicism.[94] Second, consistent with Di Tella, et al.'s, interpretation of their findings, the relative positivity of Ann's beliefs about Bob's behavior will be lower as Ann has a greater opportunity to take money from Bob. Thus, viewed jointly, these predictions suggest that, at least in many settings, strategic cynicism may be a relative rather than an absolute phenomenon.

The above analysis yields a straightforward interpretation for the absence of absolute strategic cynicism in our first experiment and the presence of relative strategic cynicism in

[94] This prediction also arises under the model that Di Tella, et al., use to motivate their experiment, though their analysis does not investigate this property.

the study by Di Tella, et al. However, existing empirical support for the above two predictions involves comparisons across studies, in which changing populations or incidental factors may yield varying results. Therefore, in a third step, we test the two predictions in a novel experiment. We conduct a replication of Di Tella, et al.'s, study, but also elicit the beliefs of neutral observers regarding the behavior of subjects in the role of Bob.[95]

In this new experiment, we replicate Di Tella, et al.'s, main finding—a comparative static result that individuals with a greater opportunity to take from their counterparts hold less positive beliefs about these opponents. This replication itself is noteworthy, as we have a substantially larger sample size and find qualitatively similar findings in a different population, in Switzerland rather than Argentina, in a society that differs in general levels of corruption, trust and trustworthiness. However, we also once more document a lack of strategic cynicism in absolute terms. The beliefs of individuals in the role of Ann with a strong incentive to engage in strategic cynicism are no more cynical about Bob's behavior than either the empirical frequency of actual choices or the beliefs of neutral third parties without any incentive to adopt a negative view of Bob's likely actions.

Our results should not be interpreted as questioning Di Tella, et al.'s, findings. In fact, we provide a direct replication of their main result of relative strategic cynicism. However, we additionally provide clear evidence—across both of our studies and in Di Tella, et al.'s, original data—that there is no strategic cynicism in absolute terms. Instead, we find that strategic beliefs are positively biased. Our contribution thus expands our understanding of the psychological forces behind self-serving belief manipulation, by noting that strategic cynicism may compete with a tendency towards positivity in determining individuals' beliefs. Such a tendency towards positivity is consistent with overwhelming evidence of a general "positivity illusion" (Taylor and Brown, 1988) from psychological studies: people hold overoptimistic beliefs about future life events (Weinstein 1980, 1989), are too optimistic about the degree of personal control (Langer, 1975), hold too positive perceptions of themselves (Svenson, 1981; Quattrone and Tversky, 1984), engage in wishful thinking (Irwin, 1953), and hold beliefs that the world is just (Lerner, 1980).[96] This positivity bias can also explain why, in contrast to our study, many other studies found strong evidence for

[95] Di Tella, et al., argue that one of their experimental treatments provides an estimate of unbiased beliefs. However, as we discuss in detail below (see footnote 107), there are a few reasons why these estimates are unlikely to correspond to the unbiased beliefs of subjects in their main experiment.

[96] In economics, Haisley and Weber (2010) document a tendency to believe that the impacts of one's choices on others are more positive than they actually are, while Andreoni and Sanchez (2014) find that subjects are too optimistic about other players' trust and trustworthiness compared to actual behavior.

motivated reasoning (in non-strategic settings). We therefore contribute to a better understanding of the specific contexts, in which we should expect biased beliefs to arise.[97]

The next section provides a detailed description of our first study using the pre-emptive taking game. In Section 3, we discuss the study by Di Tella, et al., show that their findings do not provide evidence of strategic cynicism in absolute terms and present a stylized model that can provide an interpretation of behavior in both experiments. Section 4 presents our second experiment, intended to test this model more directly and reconcile the earlier results. Finally, Section 5 concludes.

4.2 Study 1: An experimental test of strategic cynicism

We first introduce the pre-emptive taking game, which allows us to test for absolute bias in strategic beliefs. Then, we discuss the experimental implementation of the game and present our results.

4.2.1 The pre-emptive taking game

There are two players, Ann and Bob. Both players start with an endowment of 10. They play a sequential game. In Stage 1, Ann decides how much to take from Bob's initial endowment. She can take any amount, $a \in \{0, 2, 4, \ldots, 10\}$. After Stage 1, Ann's wealth equals $10 + a$, while Bob's equals $10 - a$.

In Stage 2, after observing a, Bob decides how much to take from Ann's current endowment, b, once again in increments of two. The amount that Bob can take is constrained by Bob's remaining wealth. Specifically, in order to take b units from Ann, Bob has to spend $0.5b$ from his remaining wealth and cannot spend more than the amount he has at the beginning of Stage 2. Furthermore, Bob cannot take more than Ann's wealth at the beginning of Stage 2. Thus, Bob's ability to take is given by, $b \in \{0, \ldots, \bar{b}\}$, where $\bar{b} = \min(2(10 - a), 10 + a)$. Hence, in the case in which Ann took everything in Stage 1 (i.e., $a = 10$), Bob cannot take anything in Stage 2.

After Stage 2, the game concludes. The two players' payoffs are determined as follows:

[97] Other studies that demonstrate limits in the extent to which motivated reasoning and justifications facilitate egoistic behavior are van der Weele, Kulisa, Kosfeld and Friebel (2014) who find that people do not use "moral wiggle room" (see Dana, Weber and Kuang, 2007) in the context of reciprocity and Bartling and Özdemir (2017) who find that people do not employ the "replacement logic" ("if I don't do it, someone else will") in contexts with a strong social norm.

$$\pi_A = 10 + a - b$$

$$\pi_B = 10 - a + b - 0.5b$$

As an example, suppose Ann decides to take 6 in Stage 1 such that $a = 6$. At the beginning of Stage 2, Ann has 16 and Bob has 4. In this case, in Stage 2, Bob can spend up to 4 to take up to 8 from Ann; doing so would leave both Ann and Bob with final payoffs of 8. Under standard egoistic preferences, in the unique subgame-perfect Nash equilibrium to the game both Ann and Bob take as much as they can, i.e., $a = 10$ and, consequently, $b = \overline{b} = 0$.

The key feature in the pre-emptive taking game is that Bob's ability to take from Ann is limited by how much he has left at the end of Stage 1. Ann can thus protect herself from Bob's potentially egoistic behavior by taking all his tokens. Therefore, suppose Ann wants to obtain as high a payoff as possible, but also feels obligated to be fair to Bob in the case he does not intend to take from her. In such a case, Ann may justify taking by convincing herself that Bob intends to act greedily—i.e., by engaging in strategic cynicism. The critical measure of strategic cynicism in studying this game is thus Ann's beliefs about Bob's behavior. In particular, eliciting these beliefs and comparing them to neutral and objective standards—the actual amount of taking by Bob and neutral observers' beliefs about Bob's taking—allows us to test whether they exhibit a systematic bias toward negativity.

4.2.2 Experimental design

At the beginning of each session, participants are randomly assigned to one of three roles: Player A (Ann), Player B (Bob) and Player C (neutral observer). Subjects are informed of their own role. Next, all subjects receive the same set of instructions. The instructions describe all decisions made by Player A, Player B and the neutral observer in detail and subjects are provided with a detailed table showing all the possible combinations of payoffs resulting from the two strategic players' actions.[98] After hearing the instructions read aloud, all participants answer questions about the decisions available to Players A and B and the consequences of these decisions.

For the pre-emptive taking game, Players A and B each start with an initial endowment of 10 chips, with each chip worth CHF 2 (≈ $2). In each pair, Player A selects how much to take from Player B (a). Player B's choices are elicited using the strategy method—Player B selects an amount to take (b_a) for every possible choice made by Player A. After both Player A and Player B have made their decisions, but before they learn about

[98] The instructions are available in the Appendix.

the payoffs, we elicit their beliefs concerning their counterpart's behavior. Player B guesses which value of a Player A selected ($\hat{a}^B$). Player A guesses a value of b for every possible value of a, or $\hat{b}_a^A$; at the end of the experiment one value of a is randomly selected to count for Player A's guess. Each subject earns an additional CHF 4 if they accurately guess the choice made by their opponent.[99]

Individuals in the role of Player C are not matched with any pair and are not directly affected by the choices made by any specific Player A or B. Hence, they act as neutral participants, who have no incentive to bias their beliefs about other participants' actions. For our purposes, they provide a measure of unbiased beliefs about the actions of Player As and Bs. Specifically, each neutral observer guesses the choice of a randomly selected Player A ($\hat{a}^C$) and the conditional choices of a randomly selected Player B ($\hat{b}_a^C$). Similarly to the other participants, each Player C gains CHF 4 for correctly guessing the behavior of a Player A and CHF 4 for correctly guessing one randomly selected option for a Player B.

After making all choices, participants are informed about their payoffs. They then answer several socio-demographic questions before they are paid in private.

We conducted seven sessions with between 30 and 36 participants, resulting in a total of 240 participants, 80 in each role.[100] All sessions took place at the Decision Sciences Lab (DeSciL) at the Federal Institute of Technology (ETH) in Zurich in 2015. Participants were recruited using hroot (Bock, Baetge and Nicklisch, 2014) from the joint subject pool of the University of Zurich and the ETH. The experiment was implemented using z-Tree (Fischbacher, 2007).

4.2.3 Results

On average, Player A took 8.0 tokens (std. dev. = 3.6) from Player B, with 50 of 80, or 62.5 percent, taking the full amount, $a = 10$. Figure I1 in the Appendix provides the full distribution of amounts taken.

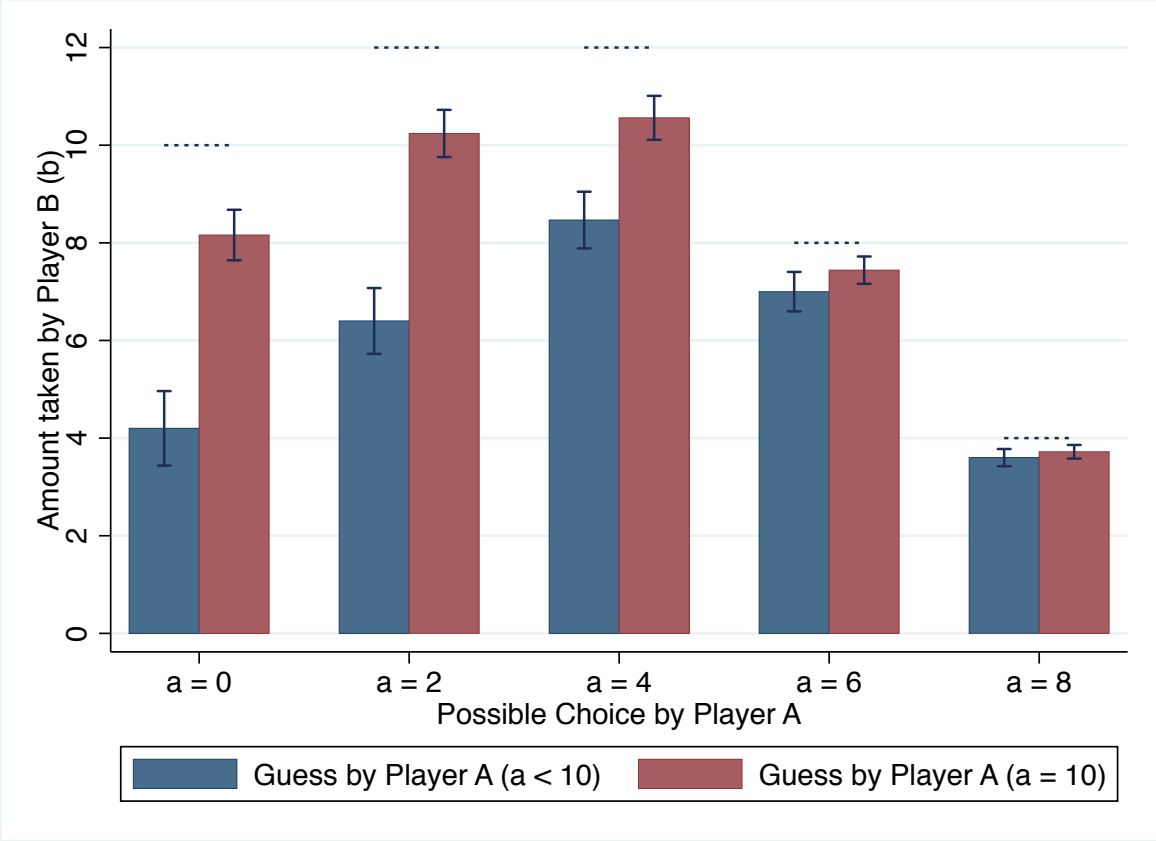
Taking by Player A is related to beliefs about how much Player B will take. Figure 16 shows the average belief of Player A regarding how much Player B will take, in response to every possible action by Player A. The figure presents these mean beliefs separately for those

[99] The amount of CHF 4 as an incentive for accurate guesses was the same for all sessions, except of the first session. In this session, the incentive for accurate guesses was CHF 2. We raised the incentive subsequently to provide subjects with more earnings opportunity. We find no differences in accuracy of guesses due to different incentives.

[100] We conducted two waves: the first four sessions were in Wave 1 while the remaining three sessions were in Wave 2. The second wave included elements intended to better ensure comprehension. We pool the data, as there is no difference in behavior between the two waves. The appendix provides instructions for both waves.

who took less than 10 ("a < 10") and those who took everything ("a = 10"). Those subjects in the role of Player A who took everything hold more cynical beliefs about Player B. For instance, for the hypothetical case where Player A takes nothing ($a = 0$), those who actually took all 10 have mean beliefs that are much more cynical (8.16) than those who took less than 10 (4.2), and this difference is highly statistically significant ($t_{78} = 4.44$, $p < 0.001$). Comparisons of mean beliefs for the cases in which Player A takes $a = 2$ or $a = 4$ similarly reveal differential cynicism between those who took 10 and those who took less (respectively, $t_{78} = 4.72$, $p < 0.001$ and $t_{78} = 2.85$, $p < 0.01$). This positive relationship between taking by Player A and negative beliefs about Player B's behavior is consistent with strategic cynicism but does not demonstrate it. Indeed, the more straightforward interpretation is that subjects in the role of Player A might simply be responding to their beliefs—taking more preemptively if they fear that B will also take more.[101]

Figure 16. Player A beliefs about Player B's actions by Player A type



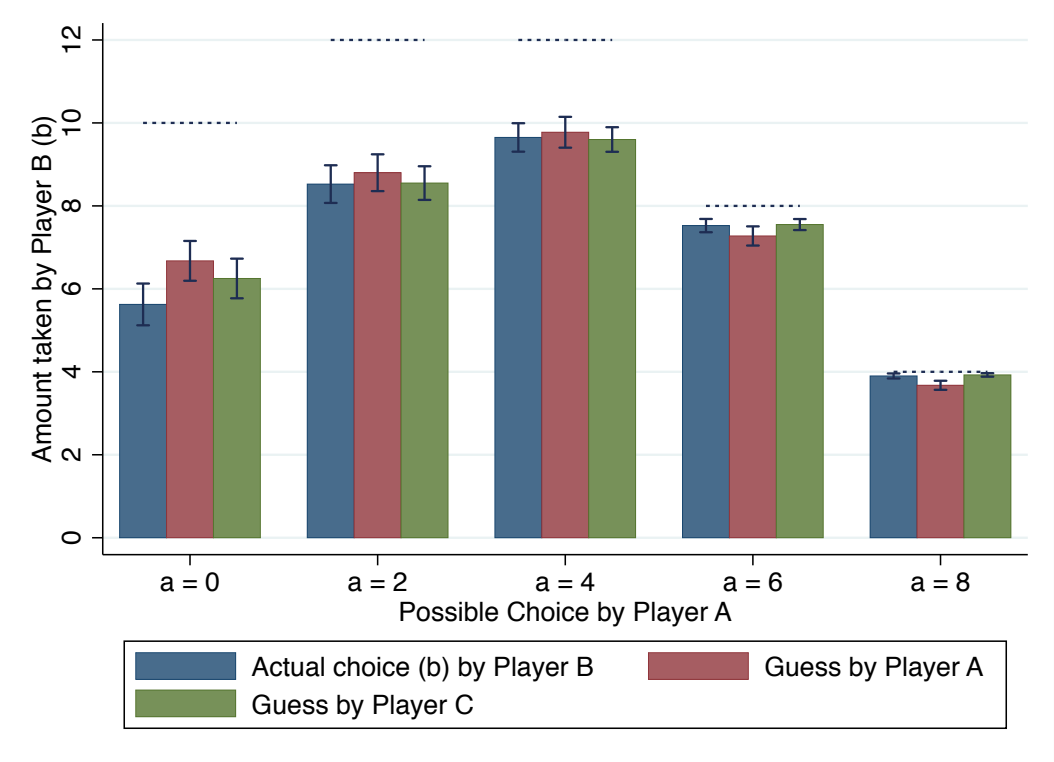
Note: The figure displays mean predictions by A of how much B will take, conditional on how much they took themselves. The category "a<10" represents Player As who took less than 10 from Player B, while the category "a=10" represents Player As who took the maximum possible from their counterparts. Dotted lines indicate the maximum possible amount B could take. Bars indicate standard errors.

[101] This is in line with results in Fischbacher and Gächter (2010) who find, using a public good game, a positive relationship between participants' own contributions and their beliefs about others' contributions.

To investigate strategic cynicism, our main focus, we next compare the beliefs of subjects in the role of Player A with two standards corresponding to unbiased beliefs. Strategic cynicism would imply that the beliefs of Player A about B's action predict more taking than the actual amount taken by Player B ($\hat{b}_a^A > b_a$) and than the corresponding predictions by Player C ($\hat{b}_a^A > \hat{b}_a^C$). Figure 17 shows, for each possible action by Player A, how much Player B actually took on average and the corresponding mean beliefs by subjects in the roles of Player A and Player C.

Looking first at the actual behavior of subjects in the role of Player B, we observe that the amount they take depends on Player A's choice. For instance, when Player A takes nothing, then the average amount taken, $b_{a=0}$, equals 5.63, even though Player B could take anywhere between 0 and 10 tokens. As Player A takes more, Player B also seizes a larger proportion of the available tokens. For instance, when Player A takes either 2 or 4 tokens, meaning that Player B can take anywhere between 0 and 12, then $b_{a=2}$ = 8.53 and $b_{a=4}$ = 9.65, on average. Finally, when A takes most of B's endowment, B takes, on average, very close to the maximum possible amounts ($b_{a=6}$ = 7.53 and $b_{a=8}$ = 3.90).[102]

Figure 17. Actual conditional taking by B and corresponding predictions by A and C



Note: Actual choices by Player Bs and predicted choices about Player B's behavior by Player As and the neutral observers (Player Cs), respectively. Dotted lines indicate the maximum possible amount B could take. Bars indicate standard errors.

[102] Figure I2 in the Appendix provides the distributions of taking by B, for every possible amount taken by A.

Perhaps the most striking finding in Figure 17 is how little evidence we find of strategic cynicism. For every possible amount taken by Player A, the beliefs of Player A and Player C regarding Player B's choice are very close to each other and to the actual behavior of Player B. Table 13 presents statistical tests of the relationships between Player A's beliefs and the actual choices by Player B and the beliefs of Player C. There are no statistically significant differences between the beliefs of Player A ($\hat{b}_a^A$) and the actual behavior of Player B (b_a) or the beliefs of Player C ($\hat{b}_a^C$) for any amount taken by A between 0 and 6. In the case where A takes 8, the differences are at least marginally statistically significant, but in these comparisons A *underestimates* B's taking, both relative to the actual amount and to the beliefs provided by C.[103]

Table 13. Statistical tests of strategic cynicism

	Mean taking by B (b_a)	Mean guess by A ($\hat{b}_a^A$)	Mean guess by C ($\hat{b}_a^C$)	$\hat{b}_a^A$ vs. b_a	$\hat{b}_a^A$ vs. $\hat{b}_a^C$
$a = 0$	5.625 (0.503)	6.675 (0.480)	6.250 (0.478)	$t_{158} = 1.510$ p = 0.133	$t_{158} = 0.678$ p = 0.531
$a = 2$	8.525 (0.454)	8.800 (0.444)	8.550 (0.406)	$t_{158} = 0.433$ p = 0.666	$t_{158} = 0.416$ p = 0.678
$a = 4$	9.650 (0.343)	9.775 (0.371)	9.600 (0.296)	$t_{158} = 0.247$ p = 0.805	$t_{158} = 0.368$ p = 0.713
$a = 6$	7.525 (0.160)	7.275 (0.231)	7.550 (0.133)	$t_{158} = 0.889$ p = 0.375	$t_{158} = 1.031$ p = 0.304
$a = 8$	3.900 (0.061)	3.675 (0.109)	3.925 (0.043)	$t_{158} = 1.800$ p = 0.074	$t_{158} = 2.130$ p = 0.035

Standard errors in parentheses

4.2.4 Discussion

To summarize, we find a positive relationship between cynicism about one's opponent and the number of tokens taken by subjects in the role of Player A. That is, subjects who hold cynical beliefs about their opponents take more from them. However, our data reveal very little evidence of self-serving belief manipulation by subjects in the role of Player

[103] The statistical tests in Table 1 are t-tests. Non-parametric Wilcoxon rank-sum tests yield very similar results; all p-values for $a = 0$ through $a = 6$ are greater than 0.188 and the p-values for $a = 8$ are 0.071 ($\hat{b}_a^A$ vs. b_a) and 0.066 ($\hat{b}_a^A$ vs. $\hat{b}_a^C$).

A, neither relative to objective behavioral standards nor to the beliefs of unbiased observers.[104]

While our results rule out significant levels of strategic cynicism in our data, they stand in contrast to Di Tella, et al. (2015), who conclude from their experimental evidence that people engage in self-serving manipulation of their strategic beliefs. We next describe this evidence and attempt to reconcile our seemingly conflicting results.

4.3 Reconciling our results with Di Tella, et al. (2015)

Di Tella, et al. (2015), study strategic cynicism using a "corruption game." In contrast with our findings, they conclude that decision makers bias their beliefs about a counterpart's egoism in response to incentives. In this section, we first describe their study and findings in detail and then offer evidence that our two sets of results are similar in that neither yields evidence of an *absolute bias* in individuals' beliefs.

4.3.1 Di Tella, et al.'s, corruption game

The study by Di Tella, et al., is based on the idea that a greater opportunity to act egoistically at the expense of a counterpart creates stronger incentives to engage in strategic cynicism. Hence, their main prediction is that those with greater opportunities to act egoistically should end up with a more pessimistic belief about the counterpart's kindness.

Di Tella, et al., test this comparative-static prediction in a "corruption game."[105] In the game, an "Allocator" and a "Seller" each start with 10 tokens. The Allocator decides how to redistribute the combined 20 tokens between herself and the Seller. Simultaneously, the Seller sets the "price" at which all the tokens are sold to the experimenter. He can either choose a price of 1.50 Argentine pesos ($) or a price of $0.50. If the Seller chooses the latter, he additionally receives a fixed side payment of $5. Setting a lower price while taking the side payment is labeled as a "corrupt" act by the Seller, akin to accepting a bribe.

In addition to deciding upon the distribution of tokens, the Allocator also provides beliefs ($\hat{p}$) about the likelihood that her paired Seller takes the corrupt action and about the share of Sellers in the experimental session that does so. The Allocator receives $5 for each

[104] The beliefs of Player Bs and Cs about A's actions are similarly unbiased (see Figure I3 in the Appendix). Recall that the mean amount taken by A is 8.0 (std. dev. = 0.358). Player Bs and C report mean beliefs that are slightly higher ($\hat{a}^B$ = 8.35 (0.280), $\hat{a}^C$ = 8.53 (0.258)) but these differences are not statistically significant—all comparisons using t-tests or rank-sum tests yield p > 0.23.

[105] Di Tella, et al., present the results of two different corruption games. We describe the game used in their preferred study, the modified corruption game. The formal games only differ in their payoffs.

correct guess. These beliefs—the key measure in Di Tella, et al.'s, experiment—provide estimates of the (possibly biased) beliefs that the Allocator has about Sellers' behavior.

Di Tella, et al., identify strategic cynicism with a treatment distinction that varies constraints on Allocators' ability to redistribute tokens. In the "Able = 2" treatment, the Allocator can move up to two tokens; that is, she can implement one of the following five payoff distributions: (8, 12), (9, 11), (10, 10), (11, 9), (12, 8). In the "Able = 8" treatment, the Allocator can move up to eight tokens, meaning that the allocations, (2, 18), (3, 17), …, (17, 3), (18, 2), are all possible. Hence, the treatment manipulation endows some Allocators with the ability to appropriate up to eight of the Sellers' tokens and other Allocators with the ability to appropriate only up to two tokens. Importantly, however, a Seller is not informed of whether his paired Allocator can move two or eight tokens, meaning that Allocators, who are aware of Sellers' ignorance regarding the treatment, should form beliefs about Sellers' behavior that are independent of the treatment.[106] Instead, Di Tella, et al., argue that the treatment manipulation affects the value of engaging in self-deception, as "allocators who can take more tokens from the seller (i.e., Able = 8 instead of Able = 2) have more incentives to convince themselves that the seller is unkind" (p. 3417), that is, $\hat{p}_{Able=2} < \hat{p}_{Able=8}$.

Di Tella, et al.'s, experimental results are consistent with this prediction. Individuals who have a greater ability to take from their counterpart take more and, more importantly, reveal more pessimistic beliefs about their counterpart's corruption. This appears to contrast with the findings of our first experiment, which reveal no strategic cynicism.

4.3.2 Relative or absolute cynicism?

To reconcile the apparent discrepancy, first note that our first study sought to identify strategic cynicism through the observation that individuals bias their beliefs about a counterpart's actions relative to the objective reality or to the beliefs of unbiased observers. Specifically, suppose there is some measure of an opponent's (expected) kindness, d—where unkindness, or $1 - d$, corresponds to b in the pre-emptive taking game or p in the corruption game. If $\hat{d}$ represents a decision maker's beliefs about the opponent's kindness, then our strategic cynicism hypothesis is that a decision maker who can take from the opponent will perceive the opponent to be less kind than he actually is, or $\hat{d} < d$. In the case of Di Tella et al.'s experiment, for instance, this corresponds to the belief that p is higher than it actually is

[106] This design leaves open the possibility that Allocators' differential beliefs are the result of the "curse of knowledge" or information projection (Camerer, Loewenstein & Weber, 1989; Madarasz, 2012), whereby decision makers find it difficult to ignore their private information when guessing others' beliefs.

(see footnote 93). Instead, our first study finds that $\hat{d} \approx d$, both when we measure d using empirical behavior as the benchmark or the beliefs of unbiased observers.

Di Tella, et al.'s, experiment and findings, however, demonstrate something different. Specifically, they test whether a decision maker with a greater incentive to take from the opponent will perceive the opponent to be less kind than will a decision maker with a reduced incentive to take. That is, if we let $\hat{d}'$ represent the beliefs of a decision maker with a restricted taking opportunity—as with Allocators in the Able = 2 condition—then Di Tella, et al.'s, findings demonstrate that those who are constrained to take less adopt relatively more positive beliefs of their opponent's kindness, $\hat{d}' > \hat{d}$. Indeed, this relative comparison is also the basis of the main theoretical proposition with which they motivate their study.

The discrepancy in our findings is straightforward to reconcile if one recognizes that a relative bias and an absolute bias may not coincide. That is, if subjects in Di Tella, et al.'s, study do not exhibit an absolute bias in the direction predicted by strategic cynicism—that is, if $\hat{d}' > \hat{d} \geq d$—then the two sets of results are entirely consistent.

Di Tella, et al., do not explicitly collect a measure of the unbiased beliefs of neutral observers.[107] However, we can compare Allocators' guesses to the empirical frequency of Sellers' corruption. In Di Tella, et al.'s, first experiment, the actual proportion of Sellers who chose the corrupt option was 75 percent. Using the more precise measure of Allocators' estimates, and the only one that was incentivized, we see that those Allocators with a greater ability to take money (Able = 8) provided estimates (69 percent) that were fairly close to the empirical benchmark and, if anything, *underestimated* Sellers' corruption. On the other hand, the Allocators who had a reduced ability to take (Able = 2) provided estimates (49 percent) that were much farther from the true percentage. Similarly, in the second experiment, the frequency of corrupt behavior by Sellers was 66 percent. The estimates provided by those who could take more (Able = 8) tended to exhibit very little bias (64 percent), while the estimates from those who could take less (Able = 2) were again biased in the direction of believing too little corruption on the part of Sellers (48 percent).

[107] Di Tella, et al., argue that an additional treatment in which Allocators are forced to take a pre-specified amount from the Seller, and in which the mean estimate is 56 percent, provides "a rough estimate of what the average [estimate] would have been in the Modified Game if Allocators had not incurred in self-deception." However, these are not the estimates of unbiased observers, but of individuals interacting with the counterpart. In the simple model we develop below, such individuals have an incentive to view the opponent kindly. In addition, eliminating the Allocator's choice altogether substantively changes the game—e.g., Sellers now confront a unilateral decision problem rather than a strategic game. Therefore, the beliefs of the Allocators in this game can only very cautiously be interpreted as corresponding to Allocators' (unbiased) beliefs in the corruption game. Finally, there are very few observations in this treatment (15 if one uses the same rules for excluding subjects as in other treatments—see footnote 25 in Di Tella, et al., 2015).

Our observations—based on our data and that of Di Tella, et al.—suggest that, to the extent a bias exists, it is one of positivity and the belief that opponents will be kinder than they actually are. While Di Tella's, et al.'s, evidence points to greater relative strategic cynicism on the part of those with more opportunity to take—i.e., $\hat{d}' > \hat{d}$—there is very little evidence of strategic cynicism on an absolute level—instead, it appears that $\hat{d}' > \hat{d} \geq d$. We next demonstrate how a very simple model can provide an interpretation for these patterns.

4.3.3 A simple model of strategic cynicism

In this section, we introduce a simple model that can account for the above patterns. We do not attempt to derive a general model of belief formation or self-deception. Instead, we study a simple and highly stylized representation of a decision in a non-strategic context, where an individual decides on a wealth allocation between herself and a counterpart and cares about this counterpart's perceived kindness or unkindness. It shares many features with the model used by Di Tella, et al., to motivate their experiment—indeed, the main predictions of our analysis can also be generated using their model.[108] We acknowledge that alternative modeling approaches may yield different predictions; however, we present this model merely as an example of the kind of model that can provide an interpretation for the above patterns we observe in our first experiment and in the study by Di Tella, et al., and that we can further test in a novel experiment.

Ann decides how to split an amount of money, normalized to 1, with Bob. Ann can take at most $K \in (0,1]$ for herself. Bob has one of two possible types, which represent the extent to which he is a "good" or "bad" person and is therefore perceived by Ann to deserve greater or less wealth: with probability, $p \in (0,1)$, Bob is a low-deservingness type (L) and with probability, $1-p$, Bob is a high-deservingness type (H). Ann is altruistic, but she cares less for the welfare of the low type. Specifically, she puts weight $d_L > 0$ on the low type's payoff and weight $d_H > d_L$ on the payoff of the high type. To incorporate motivated reasoning and self-deception, Ann can bias her belief, $\hat{p}$, about the share of low types. However, this incurs a psychological cost, $C(|p-\hat{p}|)$.[109] In addition, as in our experiments and those of Di Tella, et al., Ann is incentivized to hold unbiased beliefs by a monetary

[108] Our model differs in important points (e.g., we formally incorporate the restriction K). Di Tella, et al., discuss a result related to Proposition 2, but no result related to Proposition 1.

[109] We capture motivated reasoning similarly to Rabin (1994) and Konow (2000). For other examples of models with motivated beliefs, see Bénabou and Tirole (2006, 2011), Bénabou (2013), Brunnermeier and Parker (2008), and Bodner and Prelec (2002, 2003).

payoff function $P(|p-\hat{p}|)$. Since Ann does not know Bob's actual type, the weight she assigns to Bob's payoff equals its expected value, $E_{\hat{p}}(d) = \hat{p}d_L + (1-\hat{p})d_H$.

Ann's behavior is then captured by the following maximization problem:

$$\max_{x\in[0,K],\hat{p}\in[0,1]} \mathrm{U}(x,\hat{p}) = v(x) + E_{\hat{p}}(d)v(1-x) + P(|p-\hat{p}|) - C(|p-\hat{p}|), \quad (1)$$

where x represents the share that Ann takes for herself. The function $v\colon [0,1] \to \mathbb{R}_{>0}$ is a C^1 function with $v' > 0$, $P\colon [0,1] \to \mathbb{R}_{\geq 0}$ is a C^2 function with $P' < 0$ and $P'' \leq 0$, and $C\colon [0,1] \to \mathbb{R}_{\geq 0}$ is a C^2 function with $C' > 0$ and $C'' > 0$.[110]

This model generates two predictions that are consistent with the above empirical observations (all proofs are in Appendix J). First, Proposition 1 shows that Ann is, if anything, too positive regarding Bob's deservingness.

Proposition 1: For any $(x,\hat{p})$ *in* $argmax_{x\in[0,K],\hat{p}\in[0,1]} \mathrm{U}(x,\hat{p})$, $\hat{p} \leq p$.

This result is due to the fact that her utility increases in Bob's deservingness; that is, Ann prefers to think of Bob as being the more deserving type. Thus, in absolute terms, Ann's bias will always be to think of Bob as kinder than he actually is.

Proposition 2 states that, in relative terms, Ann will be more cynical—or less positive—when she has the opportunity to take more from Bob.

Proposition 2: Take K, K' *in* $(0,1]$ *with* $K' < K$ *and suppose that there is a unique solution to (1) for both* K *and* K'*, then* $\hat{p}' \leq \hat{p}$.

Thus, Proposition 2 provides a basis for the differences in relative beliefs between subjects who could take varying amounts from their opponent in Di Tella, et al.'s, experiment. In summary, this model predicts a (motivated) bias in belief formation that provides a basis for the absence of absolute levels of cynicism: as Ann shares a positive amount with Bob, she likes to think of him as a deserving type and this motivation is stronger as she is constrained to take less money from him.

Next, we illustrate the choice problem for a neutral observer with no stake in the outcome of game. This individual reports his beliefs, $\hat{p}^N$, about the deservingness of a random person in the role of Bob and faces incentives for accuracy: he receives a payoff

[110] Note that there exists a solution to (1) as $\mathrm{U}(x,\hat{p})$ is continuous and the feasible set is compact.

$P(|p - \hat{p}^N|)$. The observer can also adopt biased beliefs about p, but faces the cost, $C(|p - \hat{p}^N|)$, for doing so. The observer thus solves the following maximization problem:

$$\max_{\hat{p}^N \in [0,1]} U(\hat{p}^N) = P(|p - \hat{p}^N|) - C(|p - \hat{p}^N|) \quad (2)$$

where again $C' > 0$ and $P' < 0$. The observer maximizes his utility by having accurate beliefs:

Proposition 3: $\hat{p}^N = p$ *is the unique solution to* $\max_{\hat{p}^N \in [0,1]} U(\hat{p}^N)$.

Therefore, neutral observers report beliefs that correspond to the unbiased beliefs p.[111]

Note that this simple model can account for the pattern observed in the above laboratory results. Our results in Study 1 are consistent with Propositions 2 and 3: $\hat{p} \leq p = \hat{p}^N$. In the case of Di Tella, et al.'s, experiment, if we allow Able = 2 to correspond to K' and Able = 8 to correspond to K, the patterns in the data are consistent with both Propositions 1 and 2: $\hat{p}' \leq \hat{p} \leq p$. We next test these predictions jointly—specifically, that $\hat{p}' \leq \hat{p} \leq \hat{p}^N = p$—in a novel experiment.

4.4 Study 2: Jointly testing absolute and relative strategic cynicism

As we state above, the data from our first experiment and that of Di Tella, et al., provide, separately, support for all three of the above propositions. However, since we developed the model as a way to account for these observations, this is not particularly surprising. We next report a novel study that jointly tests all three predictions. The new experiment replicates Di Tella, et al.'s, experiment and further elicits the beliefs of neutral observers.

4.4.1 Experimental design

We began with a replication of Di Tella, et al.'s, corruption game, conducted in Switzerland. We used their instructions and replaced the monetary payoff of 1 Argentine peso with *1.20* Swiss Francs (CHF).[112] This meant, for instance, that the payment for each

[111] See Konow (2000) for a very similar result, in the case of "Benevolent Dictators."
[112] Note that we substantially increased real incentives; in 2016, CHF 1 corresponded to 7.48 Argentine pesos (PPP adjusted; OECD, 2018).

correct guess by an Allocator was *CHF 6*. In addition, we paid a participation fee of *CHF 15*.[113] We made two further substantive changes to bring their classroom experiment into a laboratory setting. First, while their experiment was fully paper based, we implemented it via computers, using the z-Tree program (Fischbacher, 2007). Second, we used a slightly different procedure to guarantee participants' anonymity.[114] We conducted nine sessions, each with between 22 and 24 participants, resulting in a total of 212 participants (106 Allocators and 106 Sellers).

We also conducted an additional variant of the experiment—which we refer to as the "neutral" treatment—to elicit the unbiased beliefs of neutral observers. In these sessions, participants received "instructions provided to a participant in a previous experiment." Specifically, each participant saw either the instructions given to an *Able = 2* or an *Able = 8* Allocator, determined at random. Since the Allocators' instructions include the instructions given to Sellers, participants also read the Sellers' instructions and had knowledge of the entire game. We made explicit to participants that, first, these were not their instructions and that, second, at the end of the experiment they could earn money by providing accurate guesses about something that happened in the previous experiment. Therefore, participants had no incentives to engage in self-deception, but still had incentives to closely attend to the instructions and understand the corruption game.

After reading the instructions, participants in the neutral treatment first had to answer the same comprehension questions as in the design of Di Tella, et al., and in our replication, to make sure that they understood the game. Subsequently, they made two guesses identical to those made by Allocators in Di Tella, et al.'s, design and in our replication: they guessed the choice made by a randomly chosen Seller in the previous experiment and they guessed what percentage of Sellers in a previous session chose the corrupt option. As in the replication, they received *CHF 6* for correct guesses, as well as a *CHF 15* participation fee. We conducted two such neutral sessions, with a total of 55 participants.

All sessions took place at the Decision Sciences Laboratory at the Federal Institute of Technology (ETH) in Zurich, in 2016. Participants were recruited using hroot (Bock, Baetge

[113] Our first session paid only a participation fee of CHF 10. We adjusted this payment upward to reflect the longer duration of the study than we originally expected.

[114] At the beginning of each session, one participant was randomly selected to be the "monitor." The remaining participants each received a random ID number hidden in an envelope, so that the experimenter could not match the ID to the participant. Subjects entered their ID numbers in their respective computer terminals. At the end of the study, we placed the amount of money earned by each participant in an envelope labeled only with the anonymous ID number and placed all envelopes on a table that participants passed on their way out of the laboratory. The monitor, who did not know the amount contained in any of the envelopes, controlled that each participant took only the correct envelope.

and Nicklisch, 2014) from the joint subject pool of the University of Zurich and the ETH. All instructions for the replication and the neutral treatment are in the Appendix.

4.4.2 Results

Table 14 compares the behavior and guesses of Allocators in Di Tella, et al.'s, experiment and in our replication.[115] The mean tokens taken (*Tokens Taken*) by the Allocator is slightly lower in our replication. More importantly, the share of Allocators who think that their paired Seller chooses the "corrupt" side payment (*Is Corrupt*) differs significantly between the *Able = 2* and *Able = 8* treatment groups in both the original experiment and our replication. The same holds for the average stated belief of Allocators regarding the share of Sellers who choose the side payment (*%-Corrupt*). Thus, despite the differences between the two studies—e.g., lab vs. field, Switzerland vs. Argentina—we replicate Di Tella, et al.'s, findings of relative strategic cynicism ($\hat{p}' \leq \hat{p}$).[116]

Table 14. Allocator behavior in Di Tella, et al., and our replication

	Modified Game (Di Tella, et al.)			*Replication*		
	Able = 2	*Able = 8*	*p-value*	*Able = 2*	*Able = 8*	*p-value*
Tokens Taken	1.35 (0.16)	6.59 (0.33)	<0.01	0.98 (0.14)	4.79 (0.44)	<0.01
Is Corrupt	0.48 (0.09)	0.85 (0.06)	<0.01	0.19 (0.05)	0.49 (0.07)	<0.01
%-Corrupt	0.48 (0.04)	0.64 (0.03)	<0.01	0.33 (0.04)	0.47 (0.04)	<0.01
N	31	34		53	53	

Notes: Standard errors in parentheses. P-value: t-test of the null hypothesis that the means under Able = 2 and Able = 8 are equal.

We next compare the beliefs provided by Allocators with those of neutral observers, to see whether $\hat{p}' \leq \hat{p} \leq \hat{p}^N$ holds in our new study. Table 15 compares the estimates of unbiased beliefs we obtained in our neutral treatment with the estimates provided by both

[115] Following Di Tella, et al., we conduct randomization tests with respect to demographic measures (gender, age and socioeconomic class). We find no significant differences between the Able = 2, Able = 8 and Neutral treatments.

[116] The statistical tests in Table 2 are t-tests. Non-parametric Wilcoxon rank-sum tests yield very similar results; p-values for Tokens Taken, Is Corrupt and %-Corrupt are smaller than 0.01.

Able = 2 and *Able = 8* Allocators in our replication. The final two columns report statistical tests of the differences between neutral observers' beliefs and those of the two types of Allocators. The neutral and unbiased estimate is 44 percent for *Is Corrupt* and 46 percent for *%-Corrupt*.[117] Both of these are close to—and statistically indistinguishable from—the beliefs provided by Allocators with the high opportunity to take (Able = 8), again suggesting that these allocators exhibit no bias (i.e., $\hat{p} = \hat{p}^N$). Moreover, the estimates of 44 and 46 percent are very close to the actual frequency of corrupt choices by Sellers (42 percent). Thus, similarly to our first experiment, neutral observers provide fairly accurate estimates of behavior. In contrast, the mean beliefs provided by Allocators in the *Able = 2* treatment are considerably more positive than the beliefs provided by neutral observers and these differences are statistically significant.[118]

Table 15. Allocator and neutral observer beliefs

	Replication		*Neutral Treatment*		
	Able = 2	*Able = 8*	*Neutral*	*p-value vs. Able = 2*	*p-value vs. Able = 8*
Is Corrupt	0.19 (0.05)	0.49 (0.07)	0.44 (0.07)	0.005	0.576
%-Corrupt	0.33 (0.04)	0.47 (0.04)	0.46 (0.04)	0.019	0.743
N	53	53	55		

Notes: Standard errors in parentheses. P-value: t-test of the null hypothesis that the means under neutral and the corresponding replication (Able = 2 or Able = 8) mean are equal.

The results of this study provide support for our interpretation of a key distinction between relative versus absolute strategic cynicism in our study and in the one of Di Tella, et al., and for the theoretical model we used to account for these observations. Comparing the behavior of Allocators with a high and low taking opportunity, we observe that the former hold *relatively* more cynical beliefs—i.e., that $\hat{p}' \leq \hat{p}$, as in Proposition 2. We also observe

[117] There is no significant difference between the beliefs of neutral subjects who received the instructions of an Able=2 Allocator and those who received instructions of an Able=8 Allocator.

[118] Non-parametric Wilcoxon rank-sum tests yield very similar results; p-values for Is Corrupt are 0.006 for the comparison between the neutral and the Able=2 Treatment and 0.574 for the comparison between the neutral and the Able=8 Treatment; p-values for %-Corrupt are 0.019 for the comparison between the neutral and the Able=2 Treatment and 0.835 for the comparison between the neutral and the Able=8 Treatment.

that, in absolute terms, the estimates provided by Allocators do not exhibit a tendency toward cynicism, relative to the unbiased estimates of neutral observers—i.e., $\hat{p}' \leq \hat{p} \leq \hat{p}^N$, as in Propositions 1 and 3.

4.5 Conclusion

This paper studies whether the tendency to manipulate one's beliefs self-servingly extends to strategic cynicism, whereby an individual views her opponents' likely actions negatively when doing so can justify acting in a self-interested manner. We begin with a laboratory experiment that compares the beliefs of strategic players motivated to engage in strategic cynicism with the beliefs of neutral observers not incentivized to engage in any belief manipulation. We find no evidence that strategic actors manipulate their beliefs regarding opponents' behavior, thus seemingly contradicting the hypothesis of strategic cynicism, at least in *absolute* terms.

We then attempt to reconcile this observation with the results from Di Tella, et al. (2015), who find evidence of strategic belief manipulation, whereby subjects with greater opportunity to take money from another person are more cynical about the counterpart's likely behavior. This provides evidence of *relative* strategic cynicism, meaning that individuals become comparatively more cynical about their opponents when doing so justifies more self-interested behavior. However, Di Tella, et al.'s, data reveal very little evidence of strategic cynicism in absolute terms. Thus, one possible interpretation of the apparent discrepancy is that individuals exhibit relatively more pessimistic beliefs regarding the behavior of their counterparts when they stand to gain more from doing so, but that, in absolute terms, their beliefs will tend toward positivity rather than cynicism.

We show that this interpretation is consistent with a simple model—similar to the one that Di Tella, et al., use to motivate their study—in which individuals enjoy giving more if they believe that the beneficiaries are nicer. If we allow individuals to manipulate their beliefs with some cost for doing so, then the model can explain both of the above patterns. People benefit from thinking that they are acting toward kind others, but will believe these others to be less kind when greater cynicism lowers the costs of acting self-interestedly. Admittedly, this model is not general, but it provides a useful framework for reconciling the results of the two studies.

To investigate this interpretation, we conducted a novel experimental test. Specifically, we first replicated Di Tella, et al.'s, experiment and then extended it to obtain new measures of unbiased beliefs from neutral observers. Our replication confirms Di Tella, et al.'s, observation of relative strategic cynicism. Nevertheless, we also find that any absolute bias in beliefs seems to lie in the direction of too much positivity, rather than cynicism, about counterparts' behavior, especially by those with a limited opportunity to act self-interestedly.

We thus find no evidence—across many comparisons—that decision makers justify treating another person unfairly by self-servingly adopting the belief that the counterpart herself intends to act more egoistically than is actually the case. Instead, in terms of an absolute level of bias, we find evidence for another form of motivated belief; namely, in the kinds of interactions we study here, individuals seem motivated to convince themselves of the deservingness of the counterpart, and end up with beliefs that are often too positive. The finding that people are too positive about other players' kindness supports a general tendency for distorted beliefs to lie in the direction of positivity and optimism rather than the opposite. In fact, a broad view of the literature suggests that there is little evidence that people systematically bias their beliefs in a negative direction. While it is certainly unreasonable to rule out the possibility that there are contexts in which people may also engage in such cynical self-deception—and that, in absolute terms, this may even occur in strategic settings—our analysis suggests that such a tendency is limited, at least in some contexts.

References

Akerlof, G. A., and Dickens, W. T. (1982) "The Economic Consequences of Cognitive Dissonance," American Economic Review, 72(3): 307–319.

Akerlof, G. A., and Kranton, R. E. (2000) "Economics and Identity," Quarterly Journal of Economics 115(3), 715-753.

Andreoni, J., and Sanchez, A. (2014) "Do Beliefs Justify Actions or Do Actions Justify Beliefs? An Experiment on Stated Beliefs, Revealed Beliefs, and Social-Image Motivation," working paper.

Babcock, L., and Loewenstein, G. (1997) "Explaining Bargaining Impasse: The Role of Self-Serving Biases," Journal of Economic Perspectives 11(1): 109–126.

Bartling, B., and Özdemir, Y. (2017) "The Limits to Moral Erosion in Markets: Social Norms and the Replacement Excuse," working paper.

Bénabou, R. (2013) "Groupthink: Collective Delusions in Organizations and Markets," Review of Economic Studies, 80: 429-462.

Bénabou, R., and Tirole, J. (2006) "Incentives and Prosocial Behavior," American Economic Review, 96(5): 1652-1678.

Bénabou, R., and Tirole, J. (2011). "Identity, Morals, and Taboos: Beliefs as Assets," Quarterly Journal of Economics, 126: 805-855.

Bénabou, R., and Tirole, J. (2016) "Mindful Economics: The Production, Consumption, and Value of Beliefs," Journal of Economic Perspectives, 30(3): 141-164.

Bock, O., Baetge, I., and Nicklisch, A. (2014) "hroot: Hamburg registration and organization online tool," European Economic Review, 71: 117-120.

Bodner, R., and Prelec, D. (2002) "Self-Signaling and Diagnostic Utility in Everyday Decision-Making," in Brocas, I. and Carrillo, J. eds., Collected Essays in Psychology and Economics, Oxford, UK: Oxford University Press.

Bodner, R., and Prelec, D. (2003) "Self-Signaling and Diagnostic Utility in Everyday Decision-Making," in Loewenstein, G., Read, D. and Baumeister, R.F. eds., Time and Decision, New York: Russell Sage Press.

Brunnermeier, M. K., and Parker, J. A. (2005) "Optimal Expectations," American Economic Review 95(4): 1092-1118.

Camerer, C., Loewenstein, G., and Weber, M. (1989) "The Curse of Knowledge in Economic Settings: An Experimental Analysis," Journal of Political Economy, 97(5): 1232-1254.

Chen, Z., and Gesche, T. (2017) "Persistent Bias in Advice-Giving," working paper.

Dana, J., Weber, R. A., and Kuang, J. X. (2007) "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," Economic Theory, 33(1): 67-80.

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015) "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism," American Economic Review, 105(11): 3416-42.

Eil, D., and Rao, J. M. (2011) "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," American Economic Journal: Microeconomics, 3(2): 114-138.

Exley, C. (2016) "Excusing Selfishness in Charitable Giving: The Role of Risk," Review of Economics Studies, 83(2): 587-628.

Fischbacher, U. (2007) "z-Tree: Zurich toolbox for ready-made economic experiments," Experimental Economics, 10(2): 171–178.

Fischbacher, U., Gächter, S., and Fehr, E. (2001) "Are people conditionally cooperative? Evidence from a public goods experiment," Economics letters 71(3): 397-404.

Fischbacher, U., and Gächter, S. (2010) "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods experiments," American Economic Review, 100(1): 541-556.

Gino, F., Norton, M. I., and Weber, R. A. (2016) "Motivated Bayesians: Feeling Moral While Acting Egoistically," Journal of Economic Perspectives, 30(3): 189-212.

Gneezy, U., Saccardo, S., Serra-Garcia, M., and Van Veldhuizen, R. (2018) "Bribing the Self," working paper.

Grossman, Z. J., and van der Weele, J. (2017) "Self-Image and Willful Ignorance in Social Decisions," Journal of the European Economic Association, 15(1): 173–217.

Haisley, E., and Weber, R. A. (2010) "Self-serving interpretations of ambiguity in other-regarding behavior," Games and Economic Behavior, 68(2): 634-645.

Hamman, J., Loewenstein, G., and Weber, R. A. (2010) "Self-interest through delegation: An additional rationale for the principal-agent relationship," American Economic Review, 100(4): 1826-1846.

Irwin, F. W. (1953) "Stated Expectations as Functions of Probability and Desirability of Outcomes," Journal of Personality, 21(3): 329-335.

Jecker, J., and Landy, D. (1969) "Liking a person as a function of doing him a favor," Human Relations, 22: 371-378.

Konow, J. (2000) “Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions,” American Economic Review, 90(4): 1072-1091.

Kunda, Z. (1990) “The Case for Motivated Reasoning,” Psychological Bulletin, 108(3): 480–98.

Langer, E. J. (1975) “The Illusion of Control,” Journal of Personality and Social Psychology, 32(2): 311-328.

Lerner, M. J. (1980) “The Belief in a Just World: A Fundamental Delusion.” New York: Plenum Press.

Levine, D. K. (1998) “Modeling Altruism and Spitefulness in Experiments," Review of Economic Dynamics, 1: 593-622.

Madarász, K. (2012) “Information Projection: Model and Applications,” Review of Economic Studies, 79: 961-985.

Mayraz, G. (2013) “Wishful Thinking,” working paper.

Mazar, N., Amir, O., and Ariely, D. (2008) “The Dishonesty of Honest People: A Theory of Self-Concept Maintenance,” Journal of Marketing Research, 45(6): 633-644.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2017) “Managing Self-Confidence,” working paper.

OECD (2018) “Purchasing power parities (PPP) (indicator),” Accessed on 21 August 2018.

Quatrone, G. A., and Tversky, A. (1984) “Causal versus diagnostic contingencies: On self-deception and on the voter's illusion,” Journal of Personality and Social Psychology, 46(2): 237-248.

Rabin, M. (1993) “Incorporating Fairness into Game Theory and Economics,” American Economic Review, 83(5): 1281-1302.

Rabin, M. (1994) “Cognitive dissonance and social change,” Journal of Economic Behavior and Organization, 23: 177-194.

Schopler, J., and Compere, J. S. (1971) “Effects of being kind or harsh to another on liking,” Journal of Personality and Social Psychology, 20(2): 155-159.

Schwardmann, P., and van der Weele, J. (2016) “Deception and Self-Deception,” working paper.

Shalvi, S., Gino, F., Barkan, R., and Ayal S. (2015) “Self-serving Justifications: Doing Wrong and Feeling Moral,” Current Directions in Psychological Science 24(2): 125–130.

Svenson, O. (1981) “Are we all less risky and more skillful than our fellow drivers?,” Acta Psychologica, 47: 143-148.

Taylor, S. E., and Brown, J. D. (1988) "Illusion and Well-Being: A Social Psychological Perspective on Mental Health," Psychological Bulletin, 103(2): 193-210.

Trivers, R. (2011) "The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life," New York: Basic Books.

van der Weele, J. Kulisa, J., Kosfeld, M., and Friebel, G. (2014) "Resisting Moral Wiggle Room: How Robust Is Reciprocal Behavior?," American Economic Journal: Microeconomics, 6(3): 256-264.

Weinstein, N. D. (1980) "Unrealistic optimism about future life events," Journal of Personality and Social Psychology, 39(5): 806-820.

Weinstein, N. D. (1989) "Optimistic Biases About Personal Risks," Science, 246: 1232-1233.

Zimmerman, F. (2018) "The Dynamics of Motivated Beliefs," working paper.

Appendix

Appendix A – Additional results Chapter 2

Figure A1: Labor supply for neutral and immoral work in the laboratory, last 5 periods

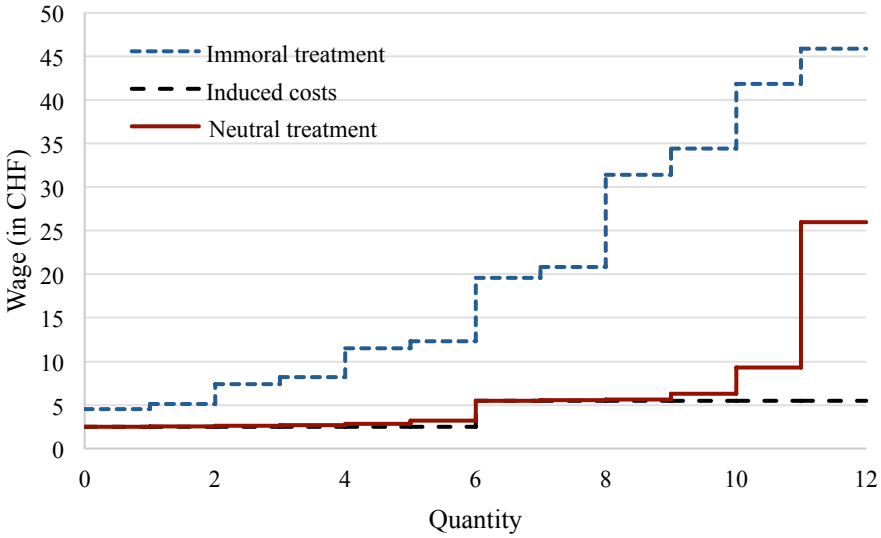
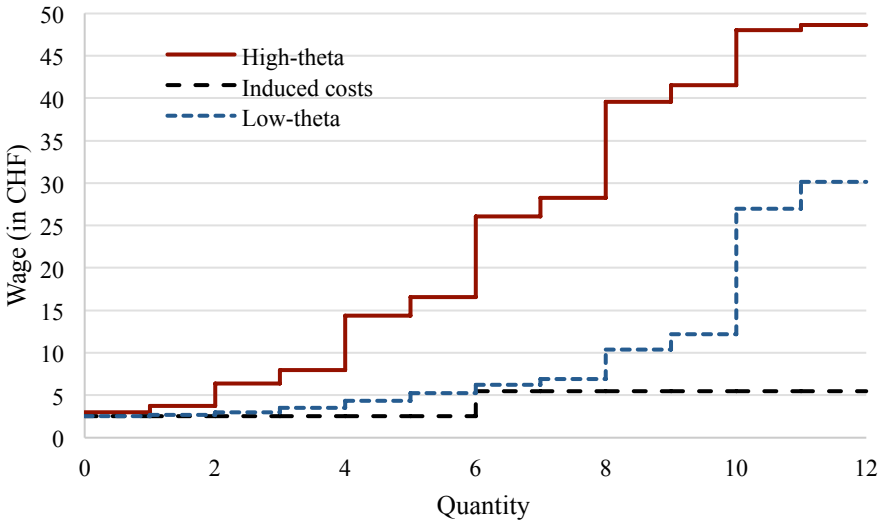


Figure A2: Labor supply for immoral work in the laboratory for different types (Θ^{Exp})



Notes: Labor supplies conditional on types are calculated with a simulation: 6 labor market decisions (first and second wage request) of high-theta (or, low-theta) types are randomly drawn (without replacement) from our sample. We then calculate the labor supply for this group of people. We repeat this 1000 times (with replacement) and take the average of these 1000 individual labor supplies. This approach differs from the one we use in Figure 4 and Figure A1, where we take the average of the actual labor supplies in the different market groups and periods. Figure 4 does not change substantially if we simulate market composition instead of using actual composition.

Figure A3: Market quantities in laboratory labor markets

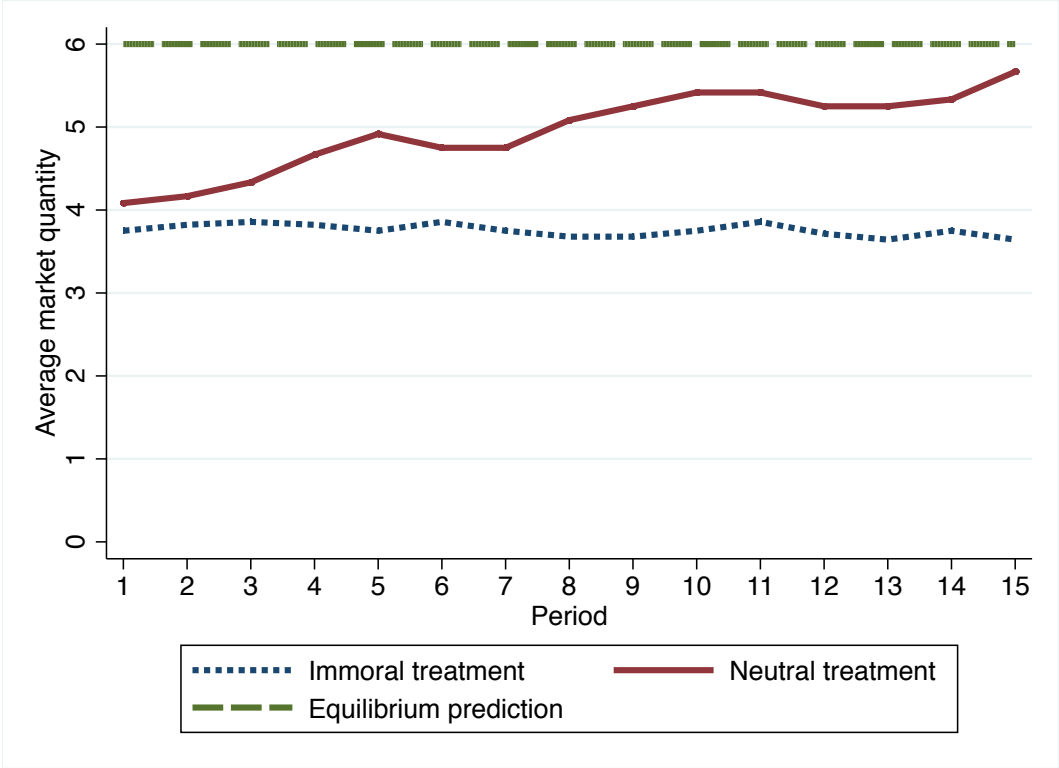


Figure A4: Employment rate by the two types in the neutral treatment

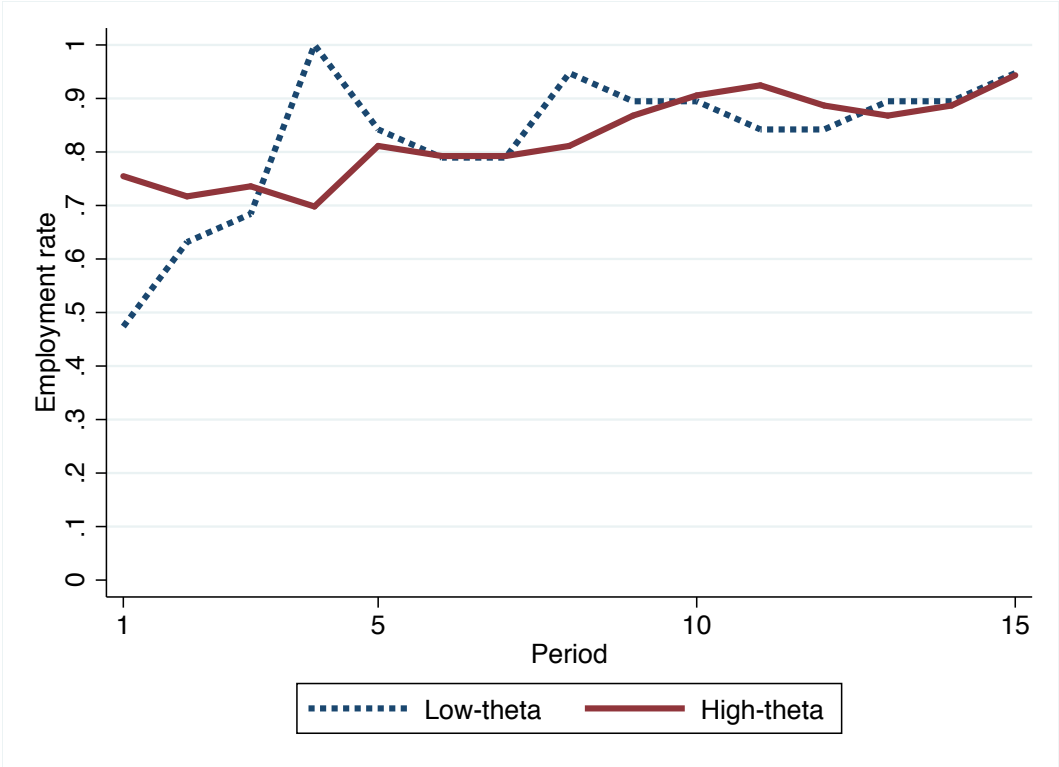


Figure A5: Income inequalities across treatments

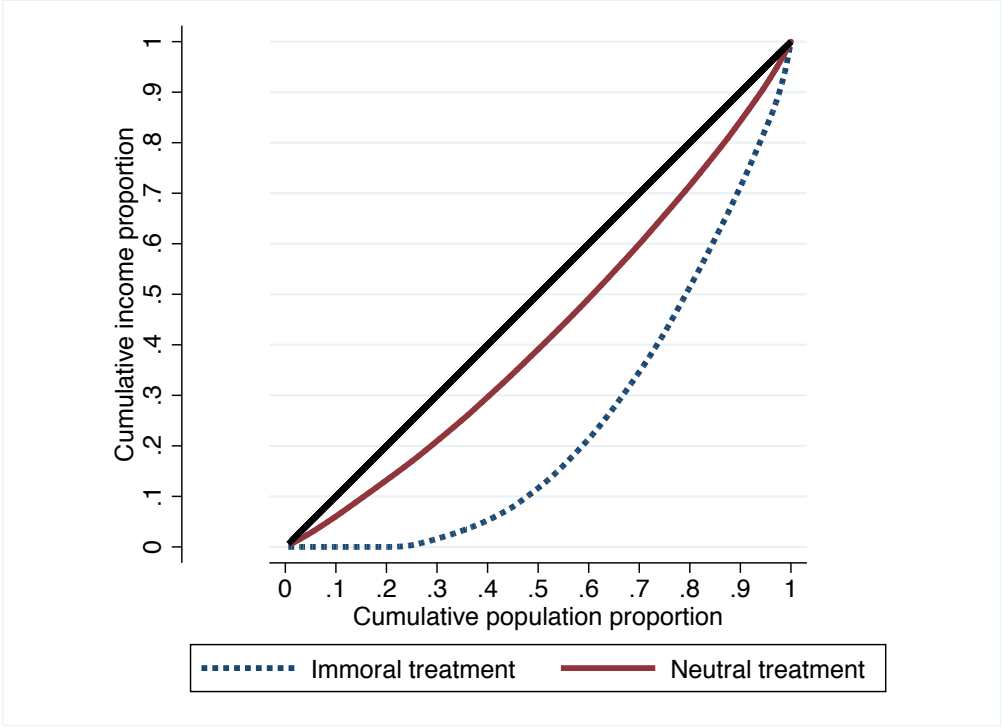


Figure A6: Probability distribution of Θ^{Sur}

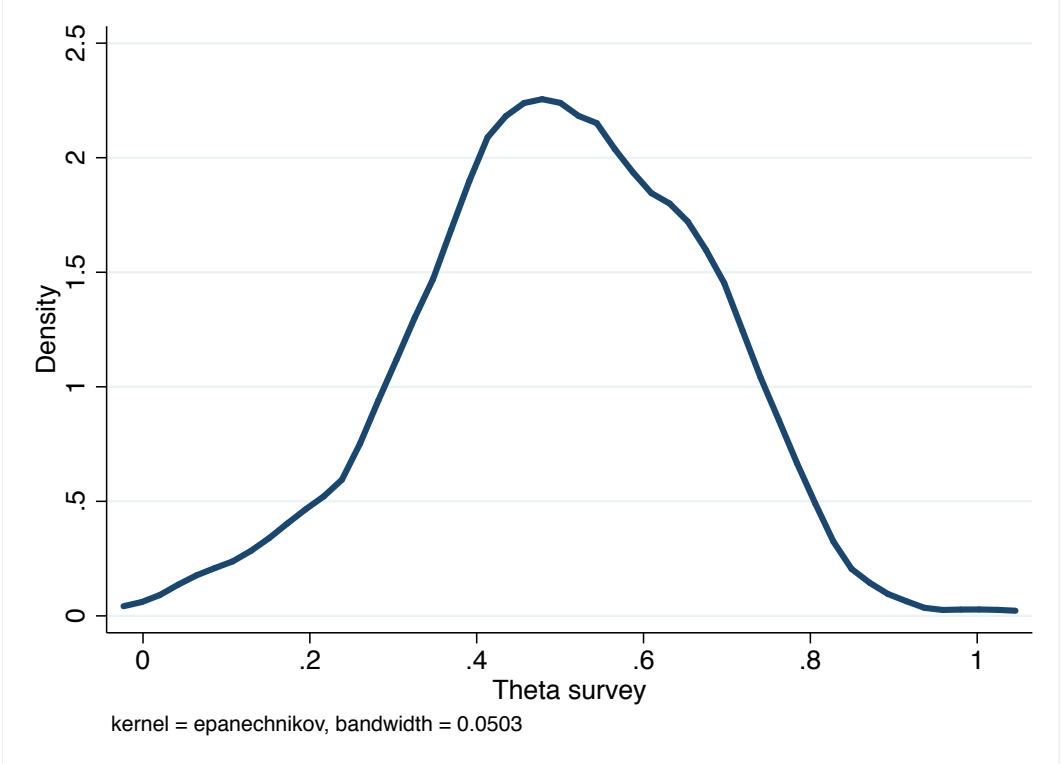
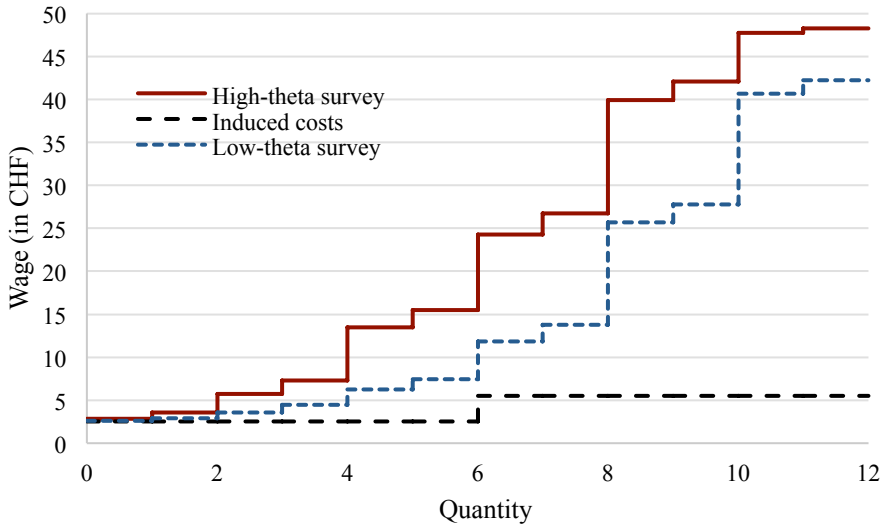


Figure A7: Labor supply for immoral work in the laboratory for different types (Θ^{Sur})



Notes: High-theta survey is 1 if θ^{Sur} is lower than the median of θ^{Sur}. Labor supplies are calculated with simulations: 6 labor market decisions (first and second wage request) of high-theta survey (or, low-theta survey) types are randomly drawn (without replacement) from our sample. We then calculate the labor supply for this group of people. We repeat this 1000 times (with replacement) and take the average of these 1000 individual labor supplies.

Table A1: Perceived immorality of industries and summary of main variables from the Swiss Labor Force Survey

Industries	N	Real gross hourly wage (in 2010 CHF)	Ln of the real gross hourly wage	Age	Male	Married	Edu high	Edu middle	Swiss	Experience	Full-time equivalent	managerial duties	Industry size	Industry Sales	Perceived immorality I(j)
All industries	32,638	41.00 (20.11)	3.59 (0.54)	42.25 (11.16)	0.57 (0.49)	0.58 (0.49)	0.35 (0.48)	0.51 (0.50)	0.64 (0.48)	9.56 (9.21)	0.87 (0.24)	0.17 (0.37)	106.46 (87.54)	0.94 (1.27)	-0.18 (0.22)
Processing of tea and coffee	138	38.88 (14.00)	3.60 (0.35)	41.29 (10.34)	0.66 (0.47)	0.67 (0.47)	0.24 (0.43)	0.59 (0.49)	0.55 (0.50)	8.78 (8.94)	0.93 (0.20)	0.10 (0.30)	3.10 (0.43)	0.22 (0.02)	-0.11 (-0.41)
Manufacture of tobacco products	162	51.01 (20.20)	3.85 (0.44)	39.42 (8.26)	0.59 (0.49)	0.59 (0.49)	0.45 (0.50)	0.41 (0.49)	0.52 (0.50)	9.63 (8.84)	0.92 (0.10)	0.03 (0.16)	2.24 (0.15)	11.79 (0.97)	0.47 (-0.42)
Manufacture of paper and paperboard	96	42.33 (16.25)	3.69 (0.33)	46.00 (10.77)	0.78 (0.42)	0.77 (0.42)	0.35 (0.48)	0.50 (0.50)	0.57 (0.50)	16.80 (13.32)	0.96 (0.10)	0.09 (0.29)	1.58 (0.22)	2.11 (1.11)	-0.06 (-0.37)
Manufacture of weapons and ammunitions	104	53.26 (20.66)	3.91 (0.37)	48.53 (10.85)	0.95 (0.22)	0.58 (0.50)	0.57 (0.50)	0.37 (0.48)	0.83 (0.38)	16.36 (13.49)	0.96 (0.06)	0.10 (0.31)	1.44 (0.10)	0.46 (0.05)	0.71 (-0.40)
Manufacture of electronic components	1,133	44.79 (18.21)	3.72 (0.41)	43.27 (10.59)	0.68 (0.47)	0.61 (0.49)	0.51 (0.50)	0.37 (0.48)	0.53 (0.50)	10.47 (9.64)	0.93 (0.14)	0.06 (0.24)	22.08 (0.76)	0.66 (0.16)	-0.01 (-0.42)
Construction of buildings	3,600	37.87 (13.18)	3.57 (0.40)	43.21 (10.73)	0.91 (0.28)	0.70 (0.46)	0.21 (0.41)	0.48 (0.50)	0.48 (0.50)	11.16 (9.89)	0.95 (0.15)	0.14 (0.34)	82.97 (0.79)	0.43 (0.01)	-0.28 (-0.39)
Maintenance and repair of motor vehicles	2,664	34.64 (12.89)	3.48 (0.38)	41.38 (12.08)	0.82 (0.38)	0.61 (0.49)	0.19 (0.40)	0.71 (0.45)	0.66 (0.47)	11.47 (10.47)	0.92 (0.21)	0.26 (0.44)	58.65 (0.33)	0.54 (0.01)	-0.28 (-0.40)
Wholesale of tobacco products	87	58.75 (27.28)	3.96 (0.48)	45.10 (9.98)	0.55 (0.50)	0.70 (0.46)	0.51 (0.50)	0.49 (0.50)	0.65 (0.48)	11.05 (7.86)	0.93 (0.15)	0.13 (0.33)	1.72 (0.07)	1.49 (0.19)	0.44 (-0.38)
Wholesale of clothing and footwear	306	38.92 (18.97)	3.54 (0.51)	40.21 (10.73)	0.31 (0.46)	0.45 (0.50)	0.38 (0.49)	0.53 (0.50)	0.52 (0.50)	7.16 (6.70)	0.84 (0.24)	0.16 (0.37)	6.08 (0.18)	1.80 (0.22)	0.10 (-0.46)
Wholesale of perfume and cosmetics	353	53.47 (28.01)	3.84 (0.54)	39.78 (10.03)	0.34 (0.47)	0.60 (0.49)	0.58 (0.49)	0.36 (0.48)	0.40 (0.49)	7.80 (6.48)	0.89 (0.19)	0.15 (0.35)	6.19 (0.26)	1.89 (0.22)	0.12 (-0.37)

See next page for the rest of the table.

Industries	N	Real gross hourly wage (2010 CHF)	Ln of the real gross hourly wage	Age	Male	Married	Edu high	Edu middle	Swiss	Experience	Full-time equivalent	Managerial duties	Industry size	Industry Sales	Perceived immorality I(j)
Wholesale of watches and jewelry	158	44.16 (19.05)	3.71 (0.40)	43.89 (10.68)	0.42 (0.49)	0.58 (0.50)	0.46 (0.50)	0.48 (0.50)	0.55 (0.50)	8.09 (8.04)	0.88 (0.19)	0.20 (0.40)	2.65 (0.17)	2.13 (0.30)	0.04 (-0.41)
Hotels and similar accommodation	2,658	27.06 (10.58)	3.22 (0.48)	40.69 (11.24)	0.40 (0.49)	0.55 (0.50)	0.19 (0.39)	0.55 (0.50)	0.40 (0.49)	7.16 (7.59)	0.85 (0.26)	0.16 (0.36)	59.16 (0.75)	0.16 (0.00)	-0.34 (-0.37)
Restaurants and mobile food activities	5,560	25.12 (9.96)	3.14 (0.47)	40.55 (11.92)	0.43 (0.49)	0.56 (0.50)	0.14 (0.34)	0.57 (0.50)	0.46 (0.50)	6.58 (7.35)	0.79 (0.29)	0.21 (0.41)	94.45 (1.03)	0.15 (0.00)	-0.33 (-0.37)
Monetary intermediations	7,286	56.00 (21.87)	3.94 (0.44)	42.00 (10.37)	0.59 (0.49)	0.55 (0.50)	0.55 (0.50)	0.42 (0.49)	0.73 (0.44)	10.65 (9.42)	0.91 (0.18)	0.19 (0.40)	109.31 (2.05)	2.26 (0.88)	0.11 (-0.40)
Credit granting	79	53.62 (22.28)	3.90 (0.39)	40.00 (8.59)	0.46 (0.50)	0.53 (0.50)	0.51 (0.50)	0.48 (0.50)	0.67 (0.47)	5.45 (5.73)	0.89 (0.17)	0.11 (0.31)	1.23 (0.03)	1.50 (0.83)	0.15 (-0.41)
Non-life insurance	2,637	49.20 (18.82)	3.82 (0.39)	42.09 (11.02)	0.49 (0.50)	0.51 (0.50)	0.51 (0.50)	0.47 (0.50)	0.80 (0.40)	9.91 (9.07)	0.89 (0.20)	0.09 (0.29)	40.52 (0.97)	1.76 (0.76)	-0.13 (-0.44)
General public administration activities	4,563	44.42 (17.24)	3.70 (0.52)	45.36 (10.74)	0.48 (0.50)	0.60 (0.49)	0.42 (0.49)	0.52 (0.50)	0.91 (0.29)	10.71 (9.53)	0.80 (0.28)	0.14 (0.35)	311.69 (6.54)	0.02 (0.00)	-0.41 (-0.36)
Gambling and betting activities	153	38.96 (16.01)	3.59 (0.40)	41.83 (10.35)	0.61 (0.49)	0.50 (0.50)	0.31 (0.46)	0.52 (0.50)	0.45 (0.50)	7.51 (5.41)	0.89 (0.20)	0.12 (0.33)	2.69 (0.06)	0.77 (0.05)	0.42 (-0.41)
Organization and operation of sport facilities for indoor and outdoor sports events	502	34.03 (15.83)	3.40 (0.66)	44.03 (12.42)	0.50 (0.50)	0.53 (0.50)	0.27 (0.44)	0.57 (0.50)	0.70 (0.46)	7.63 (8.20)	0.72 (0.34)	0.11 (0.32)	3.69 (0.13)	0.24 (0.01)	-0.49 (-0.40)
Fitness facilities	399	30.76 (16.32)	3.29 (0.56)	40.82 (11.22)	0.28 (0.45)	0.58 (0.49)	0.28 (0.45)	0.62 (0.49)	0.67 (0.47)	6.06 (6.72)	0.54 (0.34)	0.19 (0.39)	4.61 (0.38)	0.15 (0.00)	-0.35 (-0.38)

Source: Weighed data from the SLFS, years 2010-2016 (wage and demographics), STATENT, years 2011-2016 (industry size, industry sales), Value Added Tax Statistics, years 2010-2016 (industry sales) and our own survey (perceived industry immorality). Notes: N=number of observations per industry in the SLFS dataset, Male in {0, 1}, Married in {0, 1}, Education high: higher vocational education and training or university/college, Education middle: apprenticeship, full-time vocational school, matura or pedagogical training, Education low (reference category): compulsory schooling or pre-vocational education, Swiss in {0, 1}, Experience = number of years in the firm, Full-time equivalent = (working hours /42), set to 1 for working hours >= 42, managerial duties in {0, 1}, Industry size = number employees in this industry / 1000 (2010 data is not available, we substitute it with 2011 data), Industry sales = Industry sales/number employees in this industry, Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral. Standard deviations in parentheses.

Table A2: Perceived immorality of firms

Firms	Perceived immorality I(j)	Firms	Perceived immorality I(j)
Marlboro	0.54	Swisscom	-0.07
Monsanto	0.52	Firmenich	-0.09
Glencore	0.46	Winterthur Assurance	-0.1
Philip Morris	0.46	Swiss Life	-0.13
Nestlé	0.39	Swatch	-0.17
Tamoil	0.37	Adecco	-0.18
Syngenta	0.23	ABB	-0.2
UBS	0.19	Migros	-0.38
Novartis	0.18	WWF	-0.66
Credit Suisse	0.17	Pro Juventute	-0.66
Roche	0.13	Pro Natura	-0.67
Holcim	0.03	UNICEF	-0.72
Ernst and Young	-0.05	Red cross	-0.81

Source: own survey.
Notes: Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral.

Table A3: Distribution of behavior regarding the behavioral measure of concern for morality

Number of lies	Reported number for state r: 1 2 3 4 5 6	Expected payoff lying	Frequency	Share	Classification
0 (Honest)	1 2 3 4 5 6	0	161	0.671	High-theta
1	6 2 3 4 5 6	0.83	6	0.038	Low-theta
	2 2 3 4 5 6	0.17	3		
2	6 6 3 4 5 6	1.5	12	0.104	Low-theta
	1 2 3 6 6 6	0.5	2		
	1 3 3 5 5 6	0.33	1		
	1 4 4 4 5 6	0.5	1		
	5 6 3 4 5 6	1.33	2		
	6 5 3 4 5 6	1.33	1		
	3 2 3 5 5 6	0.5	1		
	3 3 3 4 5 6	0.5	5		
3	6 6 6 4 5 6	2	11	0.050	Low-theta
	4 2 3 6 6 6	1	1		
4	6 6 6 6 5 6	2.33	3	0.017	Low-theta
	6 5 5 5 5 6	1.83	1		
5	6 6 6 6 6 6	2.5	15	0.067	Low-theta
	2 3 4 5 6 6	0.83	1		
Lied in a self-harmful manner	1 2 3 4 3 3	-0.83	1	0.054	High-theta
	1 2 3 4 4 4	-0.5	1		
	1 2 3 4 5 5	-0.17	1		
	1 3 2 5 4 6	0	1		
	1 4 2 4 5 6	0.17	1		
	1 4 6 3 5 6	0.67	1		
	2 1 3 4 5 6	0	1		
	3 4 5 4 6 2	0.5	1		
	5 1 3 6 4 2	0	1		
	5 2 3 4 1 6	0	1		
	5 4 6 4 6 5	1.5	1		
	6 2 5 5 1 3	0.17	1		
	6 6 6 6 6 5	2.33	1		

Notes: Expected payoff from lying $= \frac{1}{6}\sum_{r=1}^{6}(m_{ir} - r)$*, where* m_{ir} *is the number that individual* i *reports if the actual die roll is* r*.*

Table A4: Relationship participation decision/reservation wage and θ^{Exp} (Hurdle model)

Dependent variable:	Participate	Reservation wage	Reservation wage
	(1)	(2)	(3)
Low-theta (θ_L^{Exp})	0.925***	-0.362	-0.060
	(4.64)	(-0.99)	(-0.39)
Period	-0.019**	-0.047	-0.050***
	(-2.40)	(-1.64)	(-5.74)
Period * θ_L^{Exp}	0.012	-0.015	0.005
	(1.14)	(-0.40)	(0.54)
Constant	0.449***	4.425***	3.312***
	(3.86)	(14.76)	(25.63)
Sigma		2.630***	0.571***
		(7.69)	(10.58)
Market	Immoral	Immoral	Neutral
N	2520	1755	1077
LL (pseudo)	-1424.0	-4187.4	-924.3
p-value: t + t* θ_L^{Exp} = 0	0.422	0.001	0.0000

*Notes: Estimates from Craggs double-hurdle model: (1) is a probit models; (2) and (3) are truncated linear regressions (truncated from above at 50 CHF). Models (1) and (2) use only data from the immoral markets; model (3) uses only data from the neutral markets. For neutral markets, we do not report the regression of market participation as we have only 3 incidences where a subject did not participate. Independent variables: Low-theta in {0, 1}, Period between 1 and 15. Standard errors clustered at market level; z-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Table A5: Relationship between the behavioral measures of concern for morality and outcomes in the experimental labor markets, robustness

Dependent variable:	**Employment rate**							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Number of lies								
1 lie	0.201	0.128			0.004	-0.008		
	(1.05)	(0.53)			(0.11)	(-0.12)		
2 lies	0.220**	0.182			-0.011	-0.049		
	(2.17)	(1.39)			(-0.21)	(-0.84)		
3 lies	0.398***	0.351***			-0.296***	-0.326***		
	(5.29)	(3.57)			(-12.89)	(-15.52)		
4 lies	0.392***	0.279**			0.171***	0.167***		
	(4.94)	(2.61)			(7.43)	(7.18)		
5 lies	0.286***	0.233**			0.037	0.018		
	(3.26)	(2.48)			(0.49)	(0.17)		
self-harmful lies	0.252**	0.182			-0.0516	-0.001		
	(2.53)	(1.37)			(-1.25)	(-0.04)		
Expected payoff lying			0.127***	0.105***			0.006	-0.001
			(5.04)	(3.18)			(0.18)	(-0.03)
Constant	0.475***	0.517***	0.510***	0.512***	0.829***	0.806***	0.824***	0.801***
	(11.6)	(32.97)	(13.79)	(37.01)	(36.11)	(35.35)	(35.65)	(36.06)
Market	Immoral	Immoral	Immoral	Immoral	Neutral	Neutral	Neutral	Neutral
N	168	168	168	168	72	72	72	72
R^2	0.121	0.263	0.0752	0.238	0.085	0.250	0.001	0.162
Market FE	No	Yes	No	Yes	No	Yes	No	Yes

Notes: Coefficient estimates of linear regression models. Models (1)-(4) use only data from the immoral markets, models (5)-(8) use only data from the neutral markets. Expected payoff from lying $= \frac{1}{6}\sum_{r=1}^{6} m_{ir} - r$ *(*$\in [-2.5, 2.5]$*), where* m_{ir} *is the number that individual* i *reports if the actual die roll is* r*. Standard errors clustered at market level; t-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Table A6: Relationship between θ^{Exp} and market income, depending on the behavior of other market participants.

Dependent variable:	Market income
Low-theta (θ_L^{Exp})	-3.679 (-1.05)
Many θ_H types	4.334* (2.05)
θ_L^{Exp} * Many θ_H types	16.348*** (3.53)
Constant	18.046*** (11.61)
N	168
R^2	0.089
p-value: Many θ_H types + θ_L^{Exp} * Many θ_H types=0	0.0002

*Notes: Coefficient estimates of linear regression models. Low-theta in {0, 1}, Many θ_H types: 0=number of (other) low-theta type workers is lower than 2, 1=number of (other) low-theta type workers is 2 or more. Standard errors clustered at market level; t-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Table A7: Description and summary statistics of survey scales

Variable	Number of items	Mean (Sd)	Interpretation
Protected value 1	5	0.75 (0.19)	1 = the person finds that the behavior of a banker who recommends sub-optimal assets to his clients because he has larger margins on them is: very outrageous, very blameworthy, very immoral, not at all acceptable and not at all praiseworthy.
Protected value 2	4	0.63 (0.18)	1 = the person thinks that truthfulness is a value that cannot be sacrificed.
Work ethics 1	1	0.38 (0.29)	1 = the person thinks that people are generally honest.
Work ethics 2	1	0.55 (0.36)	1 = the person thinks that calling sick to have a free day at work is really bad.
HEXACO sincerity	3	0.59 (0.19)	1 = the person is very sincere.
HEXACO fairness	3	0.68 (0.22)	1 = the person is very fair.
HEXACO greed avoidance	2	0.58 (0.23)	1 = the person is not at all greedy.
HEXACO modesty	2	0.66 (0.22)	1 = the person is very modest.
Charity attitude index	9	0.69 (0.13)	1 = the person's attitude towards charities is very positive.

Notes: Each subscale is constructed by taking averages over all items of the scale, and then normalized such that it lies between 0 and 1.

Table A8: Relationship between θ^{Sur} and outcomes in the experimental labor market

Dependent variable:	Employment rate					
	(1)	(2)	(3)	(4)	(5)	(6)
Type survey (θ^{Sur})	-0.008	0.036	0.104	0.150	0.0103	0.015
	(-0.06)	(0.22)	(0.45)	(0.75)	(0.06)	(0.06)
Immoral market (Im)	-0.046		0.013		0.133	
	(-0.37)		(0.07)		(0.59)	
θ^{Sur} * Im	-0.431	-0.366	-0.558	-0.537	-0.648*	-0.505
	(-1.64)	(-1.14)	(-1.53)	(-1.26)	(-1.69)	(-1.05)
Aggregation θ^{Sur}	Factor Analysis	Factor Analysis	Equal weight	Equal weight	Theta-Exp	Theta-Exp
N	237	237	237	237	237	237
R^2	0.137	0.294	0.132	0.291	0.133	0.293
p-value: θ^{Sur} + θ^{Sur}*Im = 0	0.0569	0.237	0.0924	0.311	0.0708	0.252
Market FE	No	Yes	No	Yes	No	Yes

*Notes: Coefficient estimates of linear regression models. Models differ in how we construct θ^{Sur} from the nine psychological survey measures. Columns (1) and (2) report our main results, using factor analysis to aggregate the psychological measures. Columns (3) and (4) give the result if equal weight is given to each measure instead. Columns (5) and (6) give the results if weights are determined by a regression of the survey measures on θ^{Exp}. Other independent variables: Immoral market is in {0, 1}, θ^{Sur} is in [0,1], where 0 means immoral and 1 means moral. Standard errors clustered at market level; t-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Table A9: Relationship between participation decision/reservation wage and θ^{Sur} (Hurdle model) in the immoral markets

Dependent variable:	Participate	Reservation wage	Participate	Reservation wage	Participate	Reservation wage
	(1)	(2)	(3)	(4)	(5)	(6)
Type survey (θ^{Sur})	-1.614**	0.432	-1.971**	0.571	-2.177**	0.373
	(-2.44)	(0.60)	(-2.08)	(0.61)	(-2.13)	(0.33)
Constant	1.326***	3.674***	1.806***	3.515***	1.841***	3.657***
	(3.98)	(10.39)	(2.99)	(6.12)	(3.01)	(5.25)
Sigma		2.679***		2.679***		2.680***
		(7.71)		(7.72)		(7.70)
Aggregation θ^{Sur}	Factor Analysis	Factor Analysis	Equal weight	Equal weight	Theta-Exp	Theta-Exp
N	2475	1711	2475	1711	2475	1711
LL (pseudo)	-1478.3	-4114.1	-1488.2	-4114.2	-1490.9	-4114.6

*Notes: Estimates from Craggs double-hurdle model: Regressions (1), (3) and (5) are probit models, regressions (2), (4) and (6) are truncated linear regressions (truncated from above at 50 CHF). Regressions differ in how θ^{Sur} is constructed from the nine psychological survey measures. Columns (1) and (2) report our main results, using factor analysis to aggregate the psychological measures. Columns (3) and (4) give the result if equal weight is given to each measure instead. Columns (5) and (6) give the results if weights are determined by a regression of the survey measures on θ^{Exp}. θ^{Sur} is in [0,1], where 0 means immoral and 1 means moral. Standard errors clustered at market level; z-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01. In the moral market, the coefficient of θ^{Sur} is not significant for any of the above specifications.*

Table A10: Regressions of willingness to work for diverse industries and firms on perceived immorality and moral types, robustness checks aggregation θ^{Sur}

Dependent variable:	Willingness to work for industry *j*				Willingness to work for firm *j*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Perceived immorality (I(j))	0.089	0.096	0.122	0.129	0.396***	0.391***	0.337***	0.327***
	(1.00)	(1.17)	(1.33)	(1.43)	(3.65)	(3.58)	(4.16)	(4.00)
Type survey (θ^{Sur})	-0.154*	-0.149*	-0.131	-0.158*	-0.200*	-0.267***	-0.243**	-0.298***
	(-1.93)	(-1.77)	(-1.49)	(-1.68)	(-1.84)	(-2.58)	(-2.21)	(-2.69)
θ^{Sur} * I(j)	-0.583***	-0.583***	-0.671***	-0.671***	-0.998***	-0.990***	-0.961***	-0.944***
	(-5.05)	(-5.06)	(-4.95)	(-4.98)	(-7.83)	(-7.60)	(-8.54)	(-8.26)
Aggregation θ^{Sur}	Equal weight	Equal weight	Theta-Exp	Theta-Exp	Equal weight	Equal weight	Theta-Exp	Theta-Exp
N	4715	4715	4715	4715	5064	5064	5064	5064
Control variables	No	Yes	No	Yes	No	Yes	No	Yes

*Notes: Coefficient estimates of linear regression models. Observations where subjects did not know the firm ("I don't know this organization") or did not fill out the questionnaire are excluded. Independent variables: Models differ in how we construct θ^{Sur} from the nine psychological survey measures. Column (1), (2), (5) and (6) give the result if equal weight is given to each measure. Column (3), (4), (7) and (8) give the results if weights are determined by a regression of the survey measures on θ_L^{Exp}. θ^{Sur} is in [0,1] where 0 means immoral and 1 means moral, willingness to work is in {0, 0.25, 0.5, 0.75, 1} where 0 means not at all willing to work, 0.5 means indifferent and 1 means really much willing to work. Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral. Control variables: age, gender, Swiss nationality, subject of study, average wage industry 2016 (SLFS; only for industries), industry size 2016 (STATENT; only for industries), industry sales 2015 (Value Added Tax Statistics; only for industries). Standard errors clustered at individual and industry/firm level (Cameron, Gelbach and Miller, 2011); z-statistics in parentheses; * p < 0.1; ** p < 0.05; *** p < 0.01.*

Table A11: Regressions of willingness to work for diverse industries and firms on perceived immorality and moral types, robustness checks classification firms immorality

Dependent variable:	Willingness to work for firm *j*			
	(1)	(2)	(3)	(4)
Perceived immorality (I_{Alt}(j))	0.141	0.138	-0.157**	-0.156**
	(1.63)	(1.59)	(-1.96)	(-1.96)
Type survey (θ^{Sur})	-0.184**	-0.222***		
	(-2.42)	(-2.99)		
θ^{Sur} * I_{Alt}(j)	-0.848***	-0.840***		
	(-9.22)	(-9.10)		
Type experiment (θ^{Exp})			-0.045*	-0.053**
			(-1.65)	(-2.04)
θ^{Exp} * I_{Alt}(j)			-0.174***	-0.175***
			(-4.06)	(-4.15)
N	5064	5064	5064	5064
Control variables	No	Yes	No	Yes

*Notes: Coefficient estimates of linear regression models. Perceived immorality is calculated different then in our main analysis: Clients that choose "I don't know this organization" are classified as giving neutral ratings. Dependent variable: Willingness to work is in {0, 0.25, 0.5, 0.75, 1} where 0 means not at all willing to work, 0.5 means indifferent and 1 means really much willing to work. Observations where subjects did not know the firm ("I don't know this organization") or did not fill out the questionnaire are excluded. Independent variables: (1) and (2) use θ^{Sur} (in [0,1]), while (3) and (4) use θ^{Exp} to classify participants, where θ^{Exp}=0 for low- theta types and θ^{Exp}=1 for high-theta types. Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral. Control variables: age, gender, Swiss nationality, subject of study. Standard errors clustered at individual and industry/firm level (Cameron, Gelbach and Miller, 2011); z-statistics in parentheses; * p < 0.1; ** p < 0.05; *** p < 0.01.*

Table A12: Regressions of willingness to work for diverse industries and firms on perceived immorality and moral types, robustness checks "I don't know this organization"

Dependent variable:	**Willingness to work for firm *j***							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Perceived immorality (I(j))	0.088	0.088	0.159***	0.159***	-0.123**	-0.123**	-0.092	-0.092
	(1.48)	(1.47)	(2.70)	(2.71)	(-2.21)	(-2.20)	(-1.49)	(-1.46)
Type survey (θ^{Sur})	-0.127**	-0.165***	-0.156	-0.204				
	(-2.01)	(-2.58)	(-1.11)	(-1.15)				
θ^{Sur} * I(j)	-0.613***	-0.613***	-0.792***	-0.792***				
	(-7.46)	(-7.39)	(-5.49)	(-5.46)				
Type experiment (θ^{Exp})					-0.034	-0.040*	-0.039	-0.056
					(-1.54)	(-1.91)	(-0.58)	(-0.90)
θ^{Exp}* I(j)					-0.131***	-0.131***	-0.158***	-0.158***
					(-4.47)	(-4.43)	(-2.98)	(-2.87)
N	6162	6162	1352	1352	6162	6162	1352	1352
Control variables	No	Yes	No	Yes	No	Yes	No	Yes

*Notes: Coefficient estimates of linear regression models. Dependent variable: Willingness to work is in {0, 0.25, 0.5, 0.75, 1} where 0 means not at all willing to work, 0.5 means indifferent and 1 means really much willing to work. Observations where subjects did not fill out the questionnaire are excluded. Columns (1), (2), (5) and (6): Observations where subjects did not know the firm ("I don't know this organization") are classified as having willingness to work of 0.5. Columns (3), (4), (7) and (8): only participants that did know all firms (N=52) are included. Independent variables: (1) - (4) use θ^{Sur} (in [0,1]), while (5) - (8) use θ^{Exp} to classify participants, where θ^{Exp}=0 for low- theta types and θ^{Exp}=1 for high-theta types. Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral. Control variables: age, gender, Swiss nationality, subject of study. Standard errors clustered at individual and industry/firm level (Cameron, Gelbach and Miller, 2011); z-statistics in parentheses; * p < 0.1; ** p < 0.05; *** p < 0.01.*

Table A13: Regressions of willingness to work for diverse industries and firms on employment rate in the immoral market

Dependent variable:	**Willingness to work for industry *j***		**Willingness to work for firm *j***	
	(1)	(2)	(5)	(6)
Perceived immorality (I(j))	-0.332***	-0.399***	-0.311***	-0.310***
	(-6.64)	(-6.74)	(-6.42)	(-6.33)
Employment rate (E)	0.029	0.044*	0.048	0.055
	(1.07)	(1.79)	(1.36)	(1.54)
E * I(j)	0.073**	0.073**	0.114**	0.114***
	(2.03)	(2.00)	(2.49)	(2.44)
N	3275	3275	3561	3561
Control variables	No	Yes	No	Yes

*Notes: Coefficient estimates of linear regression models. Sample incudes only subjects from the immoral market treatment. Dependent variable: Willingness to work is in {0, 0.25, 0.5, 0.75, 1} where 0 means not at all willing to work, 0.5 means indifferent and 1 means really much willing to work. Observations where subjects did not know the firm ("I don't know this organization") or did not fill out the questionnaire are excluded. Independent variables: Employment rate is the share of market periods in which the worker was employed. Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral. Control variables: age, gender, Swiss nationality, subject of study, average wage industry 2016 (SLFS; only for industries), industry size 2016 (STATENT; only for industries), industry sales 2015 (Value Added Tax Statistics; only for industries). Standard errors clustered at individual and industry/firm level (Cameron, Gelbach and Miller, 2011); z-statistics in parentheses; * p < 0.1; ** p < 0.05; *** p < 0.01.*

Appendix B – Robustness checks Swiss Labor Force Survey

A potential critique of our analysis of the Swiss Labor Force Survey in Section 3 is that we selected the potential immoral industries and the control industries ourselves. In this Appendix, we report results from two robustness checks that address this issue.

1) We use all industries in the Swiss Labor Force survey as control industries

In a first robustness check, we use all non-immoral industries in the Swiss Labor Force survey as control industries. We do not have a measure of the perceived industry immorality, I(j), for most industries. Instead of relying on such a measure, we define a set of industries as "immoral industries," and calculate wage premiums (or, wage discounts) for these industries, controlling for workers', jobs' and industries' characteristics. This approach is commonly used in the literature that studies stock returns for "sin industries" (e.g., Fabozzi, Ma and Oliphant, 2008; Hong and Kacperczyk, 2009; Blitz and Fabozzi, 2017; Colonnello, Curatola and Gioffré, 2019). We expect immoral industries to pay a positive wage premium, a compensating differential for the immoral nature of the work.

Set of immoral industries: We select the set of immoral industries based on the industry ratings. Most participants that rated the immorality of the industries agreed that it is immoral to work in the following four industries: manufacture of weapons and ammunitions, manufacture of tobacco products, wholesale of tobacco products and gambling and betting activities (see Table A1). These four industries are also typically considered to be "sin industries" in the literature on sin stocks. We focus on these four industries.[119] Given that we use the industry ratings to select the immoral industries, this robustness check still depends on our selection of immoral industries in Section 3. Note, however, that we use the entire dataset as control industries and therefore do not rely on our selection of control industries.

Results: Table B1 reports regressions of the natural logarithm of real gross hourly wages on the dummies for working in each of the four immoral industry, along with several additional control variables (Model 1, 2 and 3). All four immoral industries pay substantial wage premiums, in line with an immorality premium for immoral work. According to Model

[119] In Section 3, we discuss that two other industries, monetary intermediations and credit granting, might potentially be perceived as immoral. These two industries are not rated as substantially immoral (see Table A1). We therefore do not include them in the set of immoral industries. Note, however, that both industries pay substantial wage premiums. If we add dummies for working in these industries to Model 3, coefficients are 0.263 ($t=11.12$, $p<0.001$) for monetary intermediation and 0.282 ($t=20.51$, $p<0.001$) for credit granting.

3, individuals working in the immoral industries have (geometric) mean hourly earnings of in between 12 percent and 29 percent higher than people working in other industries.

2) We elicited perceived industry immorality for a second set of industries

As a second robustness check, we elicit the perceived industry immorality for a second set of industries. We then provide evidence for an immorality premium in this second set of industries. Unlike Section 3, we did not select any of the industries ourselves.

Set of industries: We created a list with all industries that had at least 50 observations in the Swiss Labor Force. This resulted in a list of 394 industries. We then asked five research assistants to select up to ten industries in which they think it is the most immoral to work and up to ten industries in which they think it is the most moral to work. They ranked the selected industries from most immoral (moral) to least immoral (moral). The research assistants were unaware of our research question. We then selected the five industries that the research assistants thought to be the most immoral and the five industries that they thought to be the most moral.[120] In addition to these ten industries, we randomly selected a set of 40 other industries that had at least 50 observations in the Swiss Labor Force Survey. Table B2 gives all selected industries.

Industry ratings: We elicited a measure of perceived immorality for the set of 50 industries. We recruited 45 participants drawn from the same subject pool from which we recruit participants for our laboratory experiment (but that did not participate in our experiment). These participants rated how immoral they think it is to work for each of the 50 industries on a 7-point Likert scale ranging from very immoral to very moral. We re-scaled the responses to lie on the -1 to 1 interval. Table B2 gives the ratings for all industries.

Results: Table B1 reports regressions of the natural logarithm of real gross hourly wages on the new collected measure of perceived industry immorality, along with several additional control variables (Model 4 to 6).[121] We find a substantial and statistically significant immorality premium. According to Model 6, individuals working in an industry as immoral as manufacture of weapons and ammunition (i.e., Perceived immorality = 0.47) have

[120] We calculated the average rank of each industry as follows. We first allocated points to each industry according to its rank: if an in industry was rated the most immoral industries, it received 10 points. The second most immoral industry received 9 points, etc. We then added up points for every industry and selected the five immoral and the five moral industries with the highest number of points.

[121] Numbers of observations differ substantially between industries. If we weight observations by industry size (instead of using survey weights), estimates for perceived immorality are similar (Model 4: 0.342, Model 5: 0.281, Model 6: 0.266) and statistically significant different from 0 (t=2.83, t=3.21,t=3.39, respectively).

(geometric) mean hourly earnings approximately 17 percent higher than people working in an industry with the median industry immorality rating (i.e., Perceived immorality = -0.14). [122]

[122] We obtain this number by doing the following calculation: $e^{0.209*0.74} - 1 \approx 0.167$.

Table B1: Relationship between wages and industry immorality, robustness

Dependent variable: ln of real gross hourly wage (in 2010 CHF)						
	(1)	(2)	(3)	(4)	(5)	(6)
Manufacture weapons and ammunitions	0.345***	0.154***	0.147***			
	(14.98)	(10.31)	(8.24)			
Manufacture tobacco	0.285***	0.237***	0.288***			
	(12.37)	(7.61)	(9.16)			
Wholesale tobacco	0.402***	0.327***	0.280***			
	(17.47)	(23.49)	(16.29)			
Gambling and betting	0.025	0.087***	0.123***			
	(1.08)	(7.26)	(9.44)			
Perceived industry immorality				0.327**	0.242***	0.209***
				(2.34)	(2.94)	(2.77)
Age		0.006***	0.006***		0.008***	0.007***
		(12.66)	(12.58)		(7.20)	(6.94)
Male		0.194***	0.195***		0.152***	0.155***
		(17.26)	(17.81)		(7.50)	(7.62)
Married		0.035***	0.039***		0.030	0.032
		(4.57)	(5.00)		(1.26)	(1.33)
Education high		0.586***	0.564***		0.602***	0.580***
		(31.24)	(32.32)		(13.69)	(12.91)
Education middle		0.247***	0.241***		0.297***	0.288***
		(17.07)	(16.99)		(6.93)	(6.60)
Swiss		0.0024	0.011		-0.011	-0.002
		(0.20)	(1.01)		(-0.63)	(-0.11)
Experience		0.004***	0.005***		0.004**	0.005***
		(5.75)	(6.24)		(2.49)	(2.77)
Full-time equivalent		-0.076***	-0.078***		-0.166**	-0.168**
		(-2.59)	(-2.73)		(-2.35)	(-2.44)
Managerial duties		-0.025	-0.020		0.063	0.066
		(-0.91)	(-0.74)		(0.85)	(0.93)
Industry sales		0.006**	0.005**		0.020	0.021
		(2.15)	(2.04)		(1.17)	(1.22)
Industry size (employees)		0.0004	0.0004*		0.001***	0.001***
		(1.59)	(1.83)		(3.80)	(4.08)
Constant	3.561***	2.832***		3.737***	2.876***	
	(154.57)	(79.63)		(52.83)	(32.09)	
N	239,313	236,625	236,625	47,935	47,935	47,935
Adjusted R^2	0.001	0.206	0.221	0.041	0.248	0.263
Year and Region FE	No	No	Yes	No	No	Yes

*Source: Weighed data from the SLFS, years 2010-2016 (wage and demographics), STATENT, years 2011-2016 (industry size, industry sales), Value Added Tax Statistics, years 2010-2016 (industry sales) and our own online-survey (perceived industry immorality). Notes: Manufacture weapons and ammunitions, Manufacture tobacco, Wholesale tobacco and Gambling and betting are binary variables where 1 means that the individual works in the respective industry. Perceived immorality is in [-1, 1] where -1 means very moral, 0 means neutral and 1 means very immoral. Control variables: Male in {0, 1}, Married in {0, 1}, Education high: higher vocational education and training or university/college, Education middle: apprenticeship, full-time vocational school, matura or pedagogical training, Education low (reference category): compulsory schooling or pre-vocational education, Swiss in {0, 1}, Experience = number of years in the firm, Full-time equivalent = (working hours /42), set to 1 for working hours >= 42, managerial duties in {0, 1}, Industry size = number employees in this industry / 1000 (2010 data is not available, we substitute it with 2011 data), Industry sales = Industry sales/number employees in this industry. Model 3 and 6 control for company region fixed effects (26 Swiss cantons) and year fixed effects (2010-2016). Standard errors clustered at the industry level, t-statistics in parentheses; * p < 0.1; ** p < 0.05; *** p < 0.01.*

Table B2: Set of industries and industry ratings

	Industry	Perceived immorality
Immoral industries	Manufacture of weapons and ammunition	0.60
	Wholesale of tobacco products	0.54
	Processing and preserving of meat (except poultry meat)	0.24
	Credit granting	0.23
	Processing and preserving of poultry meat	0.24
Moral industries	Social work activities without accommodation for the elderly and disabled	-0.46
	Residential care activities for the elderly and disabled	-0.70
	Fire service activities	-0.73
	Primary education	-0.78
	Hospital activities	-0.64
Other Industries	Manufacture of prepared meals and dishes	-0.03
	Wholesale of office machinery and equipment, except computers and computer peripheral equipment	-0.09
	Publishing of newspapers	-0.25
	Monetary intermediation (cantonal banks, commercial banks, stock exchange banks, private bankers; banks with a special field of business; regional banks; Raiffeisen banks; Foreign-controlled banks)	0.14
	Passenger rail transport	-0.36
	Support activities for crop production (preparation of fields; establishing a crop; treatment of crops; crop spraying; trimming of fruit trees and vines; transplanting of rice; thinning of beets; harvesting; pest control; provision of agricultural machinery with operators and crew)	-0.26
	Driving school	-0.21
	Security and commodity contracts brokerage	-0.01
	Printing of newspapers	-0.2
	Growing of other non-perennial crops (growing of swedes, mangolds, fodder roots, clover, alfalfa, sainfoin, fodder maize and other grasses; buckwheat; potted and bedding plants; beet seeds (excluding sugar beet seeds); seeds of forage plants and flower seeds; forage kale and similar forage products; production of cut flowers)	-0.26
	Plant propagation	-0.24
	Packaging activities (bottling of liquids; packaging of solids; security packaging of pharmaceutical preparations; labelling, stamping and imprinting; parcel-packing and gift-wrapping)	-0.03
	Manufacture of fasteners and screw machine products	-0.14
	Construction of residential and non-residential buildings	-0.15
	Manufacture of machinery for textile, apparel and leather production	0.06
	General medical practice activities	-0.58
	Activities of holding companies	0.11

See next page for the rest of the table.

	Industry	Perceived immorality
Other Industries	Retail sale of electrical household appliances in specialised stores	-0.19
	Wholesale trade of motor vehicle parts and accessories	-0.01
	Wholesale of flowers and plants	-0.24
	Dispensing chemist in specialised stores	-0.19
	Administration of financial markets	0.10
	Manufacture of other food products (soups and broths; artificial honey and caramel; perishable prepared foods; food supplements; yeast; extracts and juices; non-dairy milk and cheese substitutes; egg products; artificial concentrates)	-0.16
	Other personal service activities (astrological and spiritualists' activities; social activities; pet care services; genealogical organisations; tattooing and piercing studios; shoe shiners; porters; valet car parkers; concession operation of coin-operated personal service machines)	-0.18
	Manufacture of computers and peripheral equipment	-0.07
	Joinery installation	-0.19
	Wholesale of beverages	-0.08
	Camping grounds, recreational vehicle parks and trailer parks	-0.28
	Manufacture of optical instruments and photographic equipment	-0.21
	Renting and leasing of cars and light motor vehicles	-0.04
	Wholesale of other machinery and equipment (transport equipment except motor vehicles; production-line robots; wires and switches; other electrical material; machinery for use in trade, navigation and industry [except mining, construction, civil engineering and textile industry]; measuring instruments and equipment)	-0.10
	Non-specialised wholesale trade	-0.09
	Mixed Farming	-0.17
	Manufacture of electric domestic appliances	-0.07
	Life insurance	0.06
	Manufacture and processing of other glass, including technical glassware (laboratory, hygienic or pharmaceutical glassware; clock or watch glasses, optical glass and optical elements not optically worked; glassware used in imitation jewellery; glass insulators and glass insulating fittings; glass envelopes for lamps; glass figurines; glass paving blocks; glass in rods or tubes)	-0.17
	Wholesale of pharmaceutical goods	0.01
	Taxi operation	-0.14
	Child day-care activities	-0.69
	Manufacture of rusks and biscuits; manufacture of preserved pastry goods and cakes	-0.11

Notes: Immoral (Moral) industries are the industries that the research assistants selected as the most immoral (moral) industries. Other industries are 40 randomly selected industries that have at least 50 observations in the Swiss Labor Force Survey.

Appendix C – Welfare measure

In our model, accepting an immoral job has psychological costs, $\theta_i * I(j)$. According to Proposition 3, the least moral types benefit from an increase in immorality because the increase of the market wage exceeds the increase of the psychological costs. In the analysis in Section 4, we focused on subjects' market incomes, ignoring the psychological costs. The market mechanism that we apply allows us to measure subjects' psychological costs. In the following, we account for psychological costs and reconsider whether the predictions of our model (Proposition 3 and 4) are in line with the data.

In the experimental labor markets, subjects submit reservation wages. The differences between reservation wages and induced costs measure individuals' psychological costs (together with potential real effort costs). The difference between the market wage and the reservation wage is then a subject's benefits from market participation. Based on this reasoning, we calculate the following welfare measure for each participant *i*:

$$welfare_i = \sum_{r=1}^{15} \mathbf{1}\left(\underline{w}_1(i,r) < w(i,r)\right) * (w(i,r) - \underline{w}_1(i,r)) + \sum_{r=1}^{15} \mathbf{1}\left(\underline{w}_2(i,r) < w(i,r)\right) * (w(i,r) - \underline{w}_2(i,r))$$

where $w(i,r)$ is the market wage in round r in the market of individual i, $\underline{w}_1(i,r)$ is individual i's wage request for the doing a first job in round r, $\underline{w}_2(i,r)$ is individual i's wage request for doing the second job in round r and $\mathbf{1}\left(\underline{w}_j(i,r) < w(i,r)\right)$ measures whether individual i is hired for job j (first or second job) in round r.

Table C1, columns (1) and (2) presents evidence in support of Proposition 3: in the immoral treatment, the low-theta types have a CHF 7.05 higher welfare than the high-theta types (p=0.007). In the neutral treatment, the difference between types is very small (and even in the opposite direction).

However, this welfare measure has two potential shortcomings:

i) In the uniform-price sealed-offer auction, workers have incentives to submit their true reservation wage for the first job (Smith et al., 1982). For the second job, however, participants can have incentives to overstate their wage request. As a result, the second wage requests should be interpreted as upper bounds of

subjects' true reservation wages, and, as a result, the welfare measures should be interpreted as a lower bound of the "true" welfare.

ii) If some subjects do not understand the market mechanism, they might make errors in reporting their reservation wages.

We address this issue in Table C1, columns (3) – (6). First, we replicate the analysis for the welfare generated trough the first job only, that is $\sum_{r=1}^{15} \mathbf{1}\left(\underline{w}_1(i,r) < w(i,r)\right) * (w(i,r) - \underline{w}_1(i,r))$. This measure only relies on the first wage request, which elicitation is incentive compatible. Table C1, columns (3) and (4) provide the estimates for this second welfare measure. Again, we find support for Proposition 3. In columns (5) and (6), we look at the welfare generated in the last five market periods only, that is, $\sum_{r=11}^{15} \mathbf{1}\left(\underline{w}_1(i,r) < w(i,r)\right) * (w(i,r) - \underline{w}_1(i,r)) + \sum_{r=11}^{15} \mathbf{1}\left(\underline{w}_2(i,r) < w(i,r)\right) * (w(i,r) - \underline{w}_2(i,r))$. Subjects already participated in 10 market rounds, which gave them time to learn how the market works. We again find support for Proposition 3. (Note that the dependent variable aggregates welfare from 5 periods only; if anything, effect sizes are bigger than in the other specifications.)

We also have a second welfare measure: subjects' self-report happiness, which we elicit after the final market period. We do not find a statistically significant difference in happiness between the types in both treatments (see Table C1, columns (7) and (8)).

Table C1: Relationship between θ^{Exp} and welfare

Dependent variable:	**Welfare**		**Welfare, first job**		**Welfare, last 5 periods**		**Self-reported happiness**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Low-theta (θ_L^{Exp})	-1.07**	-0.99*	-1.07**	-0.99*	-0.22	-0.26	0.64	0.43
	(-2.42)	(-1.89)	(-2.42)	(-1.89)	(-1.45)	(-1.63)	(1.03)	(0.60)
Immoral market (Im)	8.83***		7.93***		3.02***		-0.07	
	(9.93)		(8.77)		(7.80)		(-0.25)	
θ_L^{Exp} * Im	8.12***	9.92***	8.60***	9.82***	3.04***	3.77***	-0.48	-0.20
	(3.22)	(3.57)	(3.45)	(3.56)	(3.14)	(3.79)	(-0.64)	(-0.22)
N	240	240	240	240	240	240	240	240
R^2	0.134	0.235	0.151	0.264	0.136	0.276	0.006	0.113
p-value: θ_L^{Exp} +θ_L^{Exp}*IM=0	0.007	0.002	0.004	0.002	0.005	0.001	0.719	0.670
Market FE	No	Yes	No	Yes	No	Yes	No	Yes

*Notes: Coefficient estimates of linear regression models. Independent variables: Low-theta in {0, 1}, Immoral market in {0, 1}. Standard errors clustered at market level; t-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

According to Proposition 4, the least moral types benefit from the presence of more moral types in immoral labor markets. Table C2 replicates the analysis from Section 4 for all

welfare measures. We find support for Proposition 4. However, there is no effect of the type-composition of the market on self-reported happiness.

Table C2: Relationship between θ^{Exp} and welfare, depending on the behavior of other market participants.

Dependent variable:	Welfare	Welfare, first job	Welfare, last 5 periods	Self-reported happiness
Low-theta (θ_L^{Exp})	-1.455 (-0.39)	-1.381 (-0.37)	-0.243 (-0.15)	-0.383 (-1.16)
Many θ_H types	0.480 (0.27)	-0.987 (-0.55)	-0.011 (-0.01)	-0.221 (-0.58)
θ_L^{Exp} * Many θ_H types	13.270*** (2.85)	14.082*** (3.00)	4.792** (2.34)	0.868 (1.58)
Constant	13.993*** (10.09)	3.919*** (2.81)	-5.484*** (-10.12)	4.736*** (16.66)
N	168	168	168	168
R^2	0.072	0.090	0.077	0.007
p-value: Many θ_H types + θ_L^{Exp} * Many θ_H types=0	0.0068	0.0086	0.0221	0.2969

*Notes: Coefficient estimates of linear regression models. Low-theta in {0, 1}, Many θ_H types: 0=number of (other) low-theta type workers is lower than 2, 1=number of (other) low-theta type workers is 2 or more. Standard errors clustered at market level; t-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Appendix D – Proofs Chapter 2

Lemma. *For all* $j \in J^{IM}$, $w^*(j)$ *exists, is unique and is in* $(\underline{u} + c, \infty)$. *For all* $j \in J \setminus J^{IM}$, $w^*(j) = \underline{u} + c$.

Proof. Suppose $j \in J^{IM}$. *Existence*: Define $f(w,j) = S(w,j) - D(w,j)$. Note that $f(\underline{u} + c, j) = 0 - (+) < 0$, $lim_{w\to\infty} f(w,j) = 1 - 0 = 1$ and $f(w,j)$ is continuous in w. By the intermediate value theorem there exists $w^*(j) \in (\underline{u} + c, \infty)$ such that $f(w^*(j), j) = 0$. *Uniqueness*: Follows from $f(w,j)$ being strictly increasing in w on $[0,\infty)$. Suppose $j \in J \setminus J^{IM}$. Then, $S(w,j) = \begin{cases} 0, & w < \underline{u} + c \\ [0,1], & w = \underline{u} + c \\ 1, & w > \underline{u} + c \end{cases}$. Note that for any $w < \underline{u} + c$, we have $D(w,j) > 0$ but $S(w,j) = 0$, and for any $w > \underline{u} + c$ we have $D(w,j) < 1$, but $S(w,j) = 1$. For $w = \underline{u} + c$, $S(w,j) = [0,1]$ and $D(w,j) \in [0,1]$, so $D(w,j) \in S(w,j)$.

Proposition 1. *For all* $j, j' \in J$ *with* $I(j) < I(j')$, $w^*(j) < w^*(j')$.

Proof. $w^*(j) < w^*(j')$: Suppose $I(j) = 0$. Then, $w^*(j) = \underline{u} + c$ and $w^*(j') > \underline{u} + c$ (see Lemma). Suppose $I(j) > 0$. Suppose that $w^*(j) \geq w^*(j')$. Then $S(w^*(j), j) > S(w^*(j'), j')$ and $D(w^*(j), j) \leq D(w^*(j'), j)$ because F and $-D$ are strictly increasing in w on $[0,\infty)$, $I(j) < I(j')$ and therefore $D(w, j') \geq D(w, j)$ for all w. So $S(w^*(j), j) - S(w^*(j'), j') + D(w^*(j'), j') - D(w^*(j), j) > 0$, a contradiction to the definition of $w^*(j)$ and $w^*(j')$.

Corollary. *For all* $j, j' \in J$ *with* $I(j) < I(j')$ *and* $\varepsilon > 0$, *there exists* $G \in \mathcal{F}_\theta$ *such that* $w^*(j', G) - w^*(j, G) \leq \varepsilon$.

Proof. Suppose that there exists a $G \in \mathcal{F}_\theta$ such that

$$G\left(\frac{\varepsilon}{I(j)}\right) = D(w^*(j, G), j).$$

Then, $w^*(j, G) = \underline{u} + c + \varepsilon$. The Lemma and Proposition 1 then imply that $w^*(j', G) \in [\underline{u} + c, \underline{u} + c + \varepsilon)$, and, as a result, $w^*(j', G) - w^*(j, G) \leq \varepsilon$.

To proof that such a $G \in \mathcal{F}_\theta$ exist, take any $H \in \mathcal{F}_\theta$ and construct G as follows:

$$G(\mathrm{x}) = \begin{cases} 0 & if\ \mathrm{x} < 0 \\ x\dfrac{I(j)}{\varepsilon} D(w^*(j,G),j) & if\ x \in [0, \dfrac{\varepsilon}{I(j)}] \\ D(w^*(j,G),j) + \big(1 - D(w^*(j,G),j)\big) H(x - \dfrac{\varepsilon}{I(j)}) & if\ \mathrm{x} > \dfrac{\varepsilon}{I(j)} \end{cases}$$

The assumptions on D imply that $D(w^*(j,G),j) = D(\underline{\mathrm{u}} + \mathrm{c} + \varepsilon, \mathrm{j}) \in (0,1)$. Note that G is continuous, strictly increasing on $[0,\infty)$, and with $F(0) = 0$. Therefore $G \in \mathcal{F}_\theta$.

Proposition 2. *For all $j \in J^{IM}$, worker i is hired iff $\theta_i \leq \frac{w^*(j)-\underline{u}-c}{I(j)} \equiv \underline{\theta}(j) > 0$.*

Proof. A worker accepts job j iff $u_i^{accept} = w^*(j) - c - \theta_i * I(j) \geq \underline{u} \Leftrightarrow \theta_i \leq \frac{w^*(j)-c-\underline{u}}{I(j)}$. $\underline{\theta}(j) > 0$: Follows from $w^*(j) > \underline{u} + c$ (see Lemma).

Proposition 3. For all $j, j' \in J$ with $I(j) < I(j')$, there exists $\tilde{\theta}(j,j') \in \mathbb{R}_{>0}$ such that $v_i(j', w^*(j')) > v_i(j, w^*(j))$ iff $\theta_i < \tilde{\theta}(j,j')$.

Proof. Suppose $I(j) = 0$. Then $v_i(j, w^*(j)) = v_i(j, \underline{u} + c) = \underline{u}$ (see Lemma). Note that $u_i^{accept}(j', w^*(j')) > \underline{u}$ iff $\theta_i < \frac{w^*(j')-\underline{u}-c}{I(j')} = \underline{\theta}(j')$, so $v_i\big(j', w^*(j')\big) = u_i^{accept}(j', w^*(j')) > \underline{u} = v_i(j, w^*(j))$ iff $\theta_i < \underline{\theta}(j') \equiv \tilde{\theta}(j,j')$. Note that $\tilde{\theta}(j,j') > 0$ by Proposition 2.

Suppose $I(j) > 0$. Note that $v_i\big(j', w^*(j')\big) > v_i\big(j, w^*(j)\big) = max\{u_i^{accept}(j, w^*(j)), \underline{u}\}$ if and only if:

i) $u_i^{accept}(j', w^*(j')) > u_i^{accept}(j, w^*(j))$

ii) $u_i^{accept}(j', w^*(j')) > \underline{u}$

Inequality i) holds iff $\theta_i < \frac{w^*(j')-w^*(j)}{I(j')-I(j)}$ and inequality ii) holds iff $\theta_i < \underline{\theta}(j')$. Therefore, we have $\tilde{\theta}(j,j') = min\left\{\frac{w^*(j')-w^*(j)}{I(j')-I(j)}, \underline{\theta}(j')\right\}$. Note that $\tilde{\theta}(j,j') > 0$ (see Proposition 1 and Proposition 2).

Proposition 4. *For all* $j \in J^{IM}$ *and* $F, G \in \mathcal{F}_\theta$ *with* $F(x) < G(x)$ *for all* $x > 0$*, there exists* $\hat{\theta}(j,F) > 0$ *such that* $v_i(j, w^*(j,F)) > v_i(j, w^*(j,G))$ *iff* $\theta_i < \hat{\theta}(j,F)$.

Proof. First, we will proof that $w^*(j,G) \leq w^*(j,F)$. Suppose not, then $w^*(j,G) > w^*(j,F)$. But then $S(w^*(j,G),j,G) > S(w^*(j,F),j,F)$ and $D(w^*(j,G),j) < D(w^*(j,F),j)$ because F, G and $-D$ are strictly increasing in w on $[0,\infty)$, and F first-order stochastically dominates G. But then $S(w^*(j,G),j,G) - S(w^*(j,F),j,F) + D(w^*(j,F),j) - D(w^*(j,G),j) > 0$, a contradiction to the definition of w^*.

Second, define $\hat{\theta}(j,F) = \frac{w^*(j,F) - \underline{u} - c}{I(j)} \equiv \underline{\theta}(j,F)$, and note that under this definition $w^*(j,F) - c - S(j) * \theta_i > \underline{u}$ iff $\theta_i < \hat{\theta}(j,F)$. ($\hat{\theta}(j,F) > 0$ follows from Proposition 2.)

To finish the proof, note that $v_i\big(j, w^*(j,F)\big) = w^*(j,F) - c - I(j) * \theta_i > \max\{\underline{u}, w^*(j,G) - c - I(j) * \theta_i\} = v_i\big(j, w^*(j,G)\big)$ for all $\theta_i < \hat{\theta}(j,F)$, and $v_i\big(j, w^*(j,F)\big) = \underline{u} = v_i\big(j, w^*(j,G)\big)$ for all $\theta_i \geq \hat{\theta}(j,F)$.

Appendix E – Alternative model interpretation

The results in Section 3 also apply for a context with 2 jobs, a neutral job j^N ($I(j^N) = 0$) and an immoral job $j^{IM} \in J^{IM}$ ($I(j^{IM}) > 0$). In the following, we show that, under some assumptions, labor demand and labor supply correspond to their counterparts in Section 3. Therefore, all results derived in Section 3 also hold in the context with 2 jobs.

Labor supply: Labor supply consists of an interval of workers, $i \in [0,1]$. As in Section 3, we assume that the utility of accepting job j of a worker of type i is given by:

$$u_i\big(j, w(j)\big) = w(j) - c - \theta_i * I(j),$$

where the parameter θ_i is distributed according to a distribution with cdf $F \in \mathcal{F}_\theta$. For all $F \in \mathcal{F}_\theta$, F is continuous, strictly increasing on $[0,1)$, and with $F(0) = 0$. Workers choose between the neutral and the immoral job. Note that every worker with $\theta_i \leq \frac{w(j^{IM}) - w(j^N)}{I(j^{IM})}$ chooses the immoral job. The labor supply for the immoral job is then given by $S(w, j^{IM}) = F(\frac{w}{I(j^{IM})})$, where $w = w(j^{IM}) - w(j^N)$ is the immorality premium. Note that the labor supply for the immoral job corresponds to the labor supply in Section 3 with $\underline{u} = c = 0$.

Labor demand: Labor demand consists of an interval of firms, $k \in [0,1]$. Each firm can either produce a neutral product or an immoral product. Firms that produce immoral products offer immoral jobs; firms that produce neutral products offer neutral jobs. Firms' profits are:

$$\pi_k(j, w) = a_k(j) - w(j),$$

where $a_k(j)$ measures firm k's earnings when producing good j. Firms choose to produce the immoral product if $\Delta a_k(j^{IM}) = a_k(j^{IM}) - a_k(j^N) \geq w$. $\Delta a_k(j^{IM})$ is distributed according to a distribution with cdf $G_{j^{IM}}$. An increase in immorality of the job does not decrease firms earnings,[123] that is, i) $G_j(0) = 0$ for all $j \in J^{IM}$, and ii) for all $j, j' \in J^{IM}$ with $I(j') > I(j)$, $G_{j'}(x) \leq G_j(x)$ for all $x \in \mathbb{R}$. In addition, $G_{j^{IM}}$ is continuous and strictly increasing on $[0, \infty)$. The labor demand for the immoral job is then given by $D(w, j^{IM}) = 1 - G_{j^{IM}}(w)$. Note that D is continuous in w, strictly decreasing in w on $[0, \infty)$, with $lim_{w \to \infty} D(w, j^{IM}) =$

[123] One interpretation is, for example, that $I(j)$ measures negative externalities in production. Avoiding these externalities is costly; decreasing the immorality therefore increases production costs (see also Rosen, 1986).

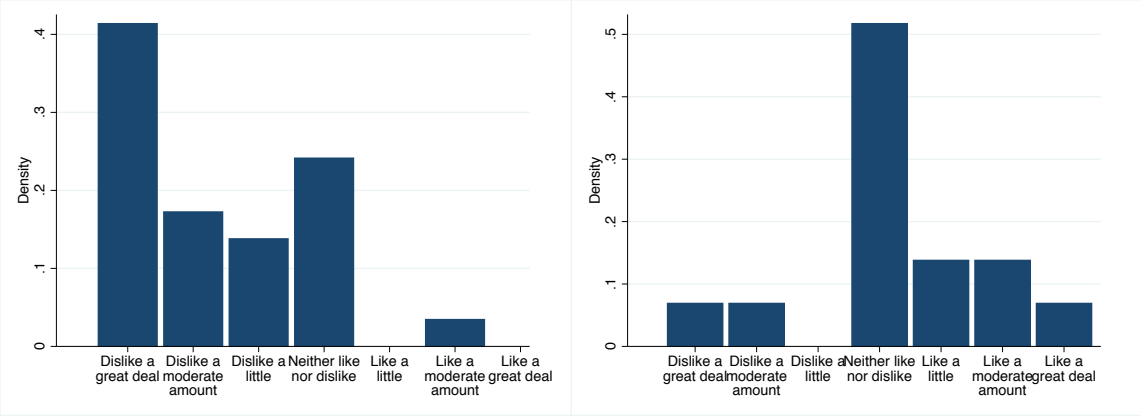
0 and $D(w, j^{IM}) = 1$ for $w \leq 0$. In addition, $I(j') > I(j)$ implies $D(w, j') \geq D(w, j)$ for all $w \in \mathbb{R}$. Note that the labor demand corresponds to the labor demand in Section 3.

The equilibrium wage, $w^*(j^{IM})$, is implicitly defined by $S(w^*(j^{IM}), j^{IM}) - D(w^*(j^{IM}), j^{IM}) = 0$.[124] As both labor demand and labor supply correspond to their counterparts in Section 3, the Lemma and Proposition 1 to 4 apply (with $j \in J^{IM}$). In particular, $w^*(j^{IM})$ is strictly positive (Lemma), so there is an immorality premium, and this immorality premium is increasing in the immorality of j^{IM}, $I(j^{IM})$ (Proposition 1). The immoral types sort into accepting the immoral jobs, while the moral types sort into accepting the neutral jobs (Proposition 2).

[124] Note that market clearance in the immoral job market implies market clearance in the neutral job market, $\big(1 - S(w^*(j^{IM}), j^{IM})\big) - \big(1 - D(w^*(j^{IM}), j^{IM})\big) = 0$.

Appendix F – Additional results Chapter 3

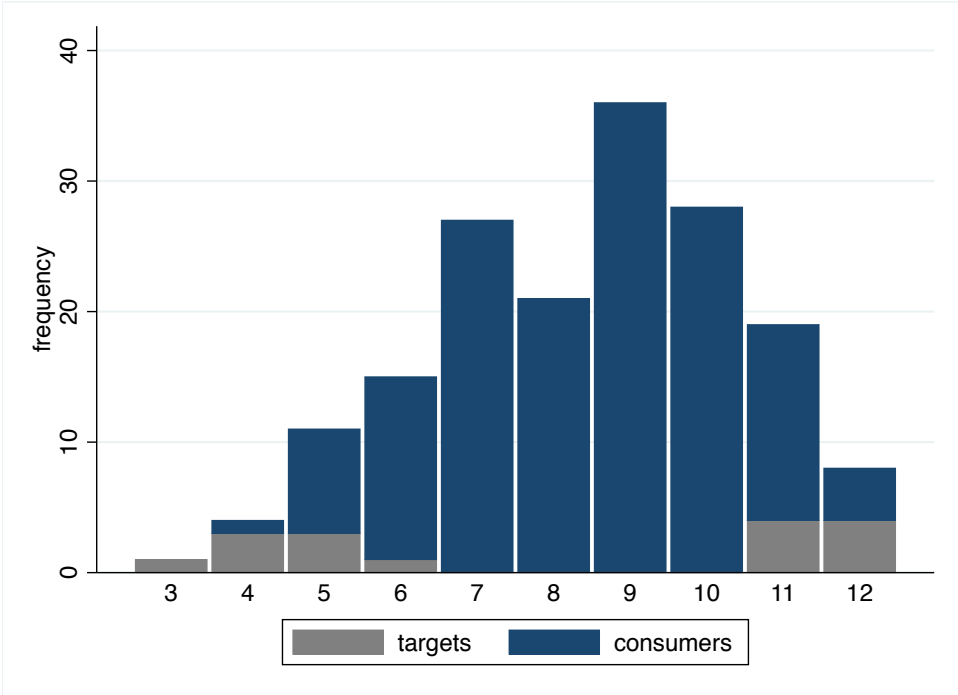
Figure F1: Desirability of public association with Zukunft CH



(a) increase the donation the most (b) reduced the donation the most

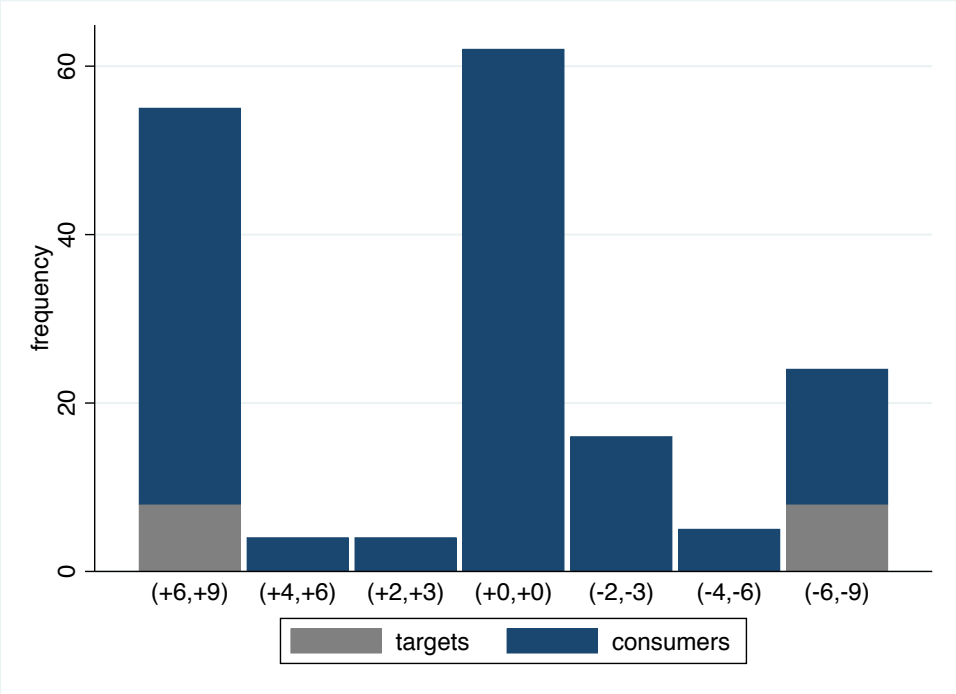
Notes: Answers to an online survey with 29 participants. Subjects are drawn from the same subjects pool that is used for Study 1. Subjects are asked to rate how much they would like it (from 1= dislike a great deal to 7= like a great deal) if other participants in a hypothetical laboratory experiment would receive information that makes them believe that "you are the participant who increased the donation to Zukunft CH the most," or information that make them believe that "you are the participant who reduced the donation to Zukunft CH the most." Figure (a) shows the distribution the first question, while Figure (b) shows the distribution to the second question. I can reject the hypothesis that both distributions are similar (Wilcoxon signed-ranks test, z=-3.702, p=0.0002).

Figure F2: Distribution intelligence score, Study 1



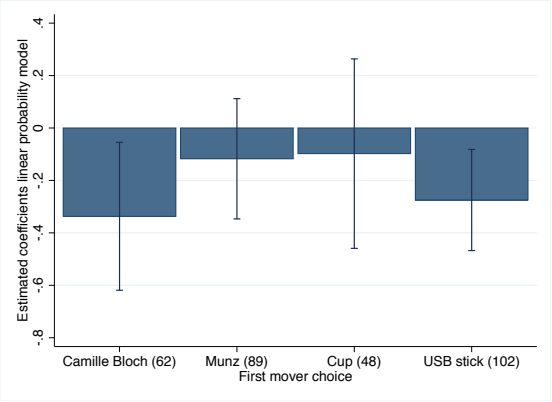
Notes: Distribution of the number of Raven's matrices correctly solved.

Figure F3: Distribution moral values, Study 1

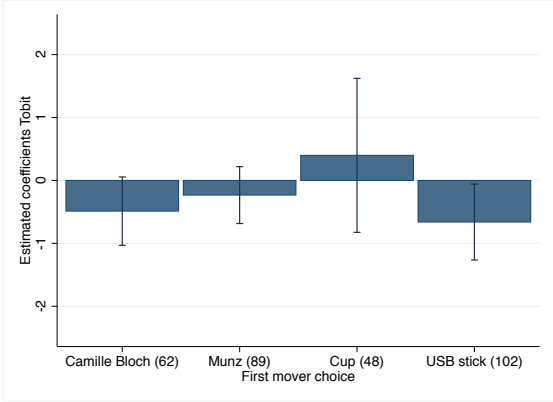


Notes: (+6, +9) means that the subjects chose the option that increases her payoff by CHF 6, and increased the donation to Zukunft CH by CHF 9.

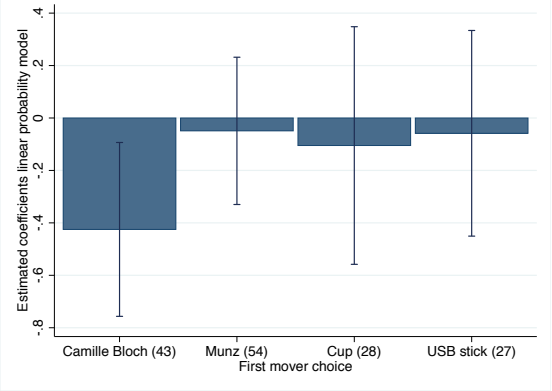
Figure F4: Treatment effects conditional on target choice, Study 1



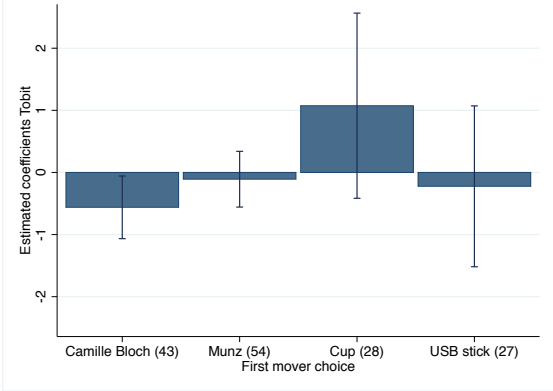
(a) pooled, probability to chose same product as the target (p joint test = 0.002)



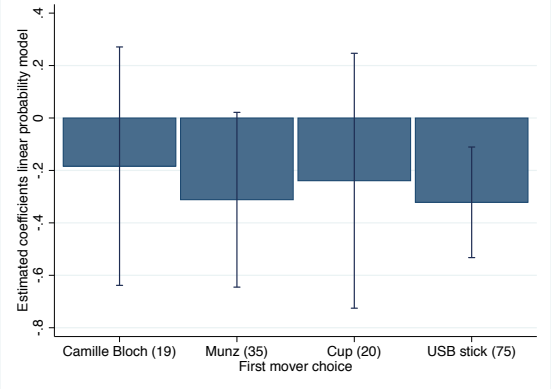
(b) pooled, WTP to receive same product as the target (p joint test = 0.011)



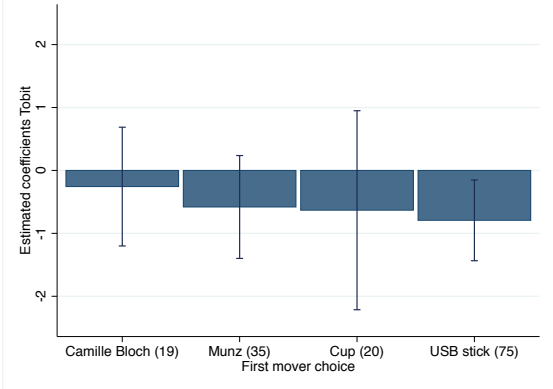
(c) intelligence, probability to chose same product as the target (p joint test = 0.128)



(d) intelligence, WTP to receive same product as the target (p joint test = 0.117)



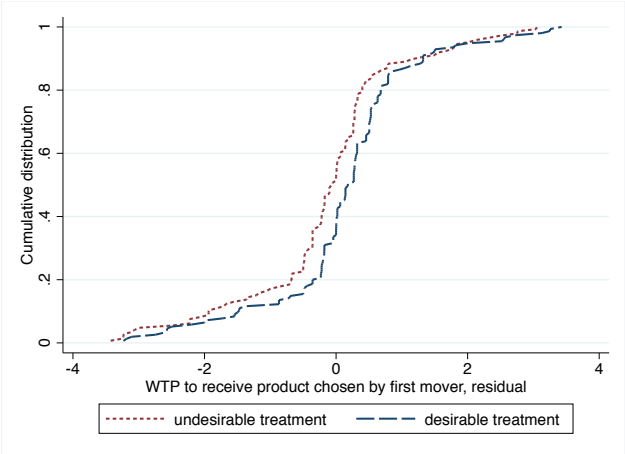
(e) moral values, probability to chose same product as the target (p joint test = 0.006)



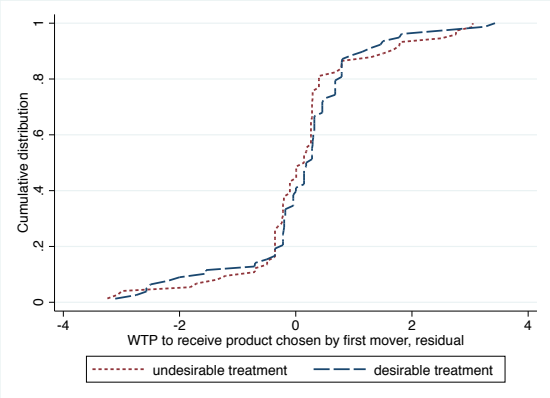
(f) moral values, WTP to receive same product as the target (p joint test = 0.044)

Notes: Bars in figure (a), (c) and (e) show estimated treatment effects from linear regressions of the probability to choose the same product as the target when none of the two products are connected with any payment on the interaction between the treatment dummy (1=undesirable treatment) and the choice of the target (1=Camille Bloch; 1=Munz; 1=Cup; 1=USB stick), "treatment x target choice." Bars in figure (b), (d) and (f) show estimated treatment effects from Tobit regressions of the willingness to pay to receive the product that was chosen by the target instead of the other product on "treatment x target choice." While figures (a) and (b) show effects if the intelligence data and moral values data is pooled, figures (c) to (f) show effects for intelligence and moral values separately. Coefficients for figure (c) and (e) and coefficients for figure (d) and (f) are estimated jointly by interacting "treatment x target choice" with an intelligence data dummy (1=intelligence), controlling for the intelligence data dummy. The numbers in brackets indicate the number of observations in this category. "p joint test" reports the p-value of a joint test that all for coefficients ("bars") are equal to zero. All regressions control for the choice of the target and session fixed effects. Bars indicate 95%-confidence intervals. Robust standard errors are used.

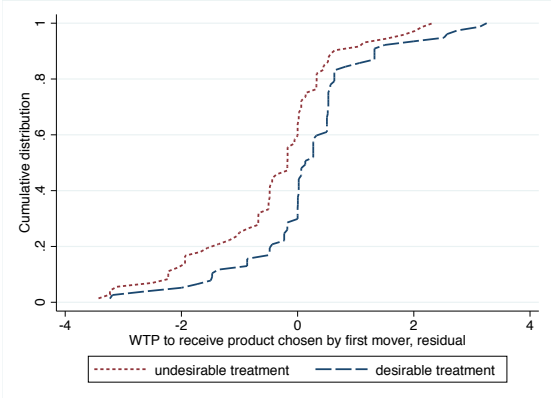
Figure F5: Distribution willingness to pay in Study 1, residuals



(a) residuals



(b) data on intelligence, residuals

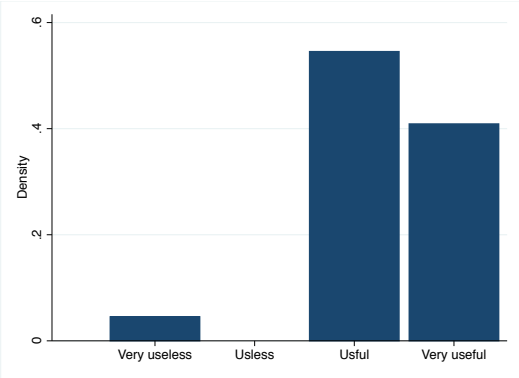


(c) data on moral values, residuals

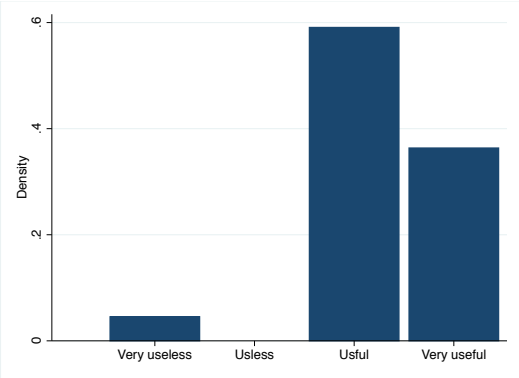
Notes: Cumulative distributions of subjects willingness to pay to receive the product chosen by the target instead of the other product. Figures control for differences in targets' choices and session fixed effects by plotting residuals from a regression of willingness to pay on targets product choice and session fixed effects. (a) uses data from both rounds, (b) uses only data from the round related to intelligence and (c) use only data from the round related to moral values.

Figure F6: Associations of chocolates with students and neo-Nazis

(a) How useful is the product for students? (Wilcoxon signed-rank test, p= 0.317, N=21)

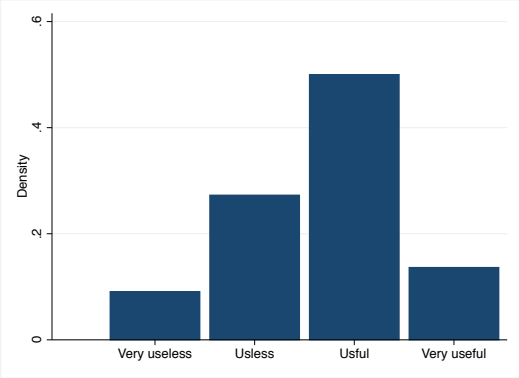


Camille Bloch

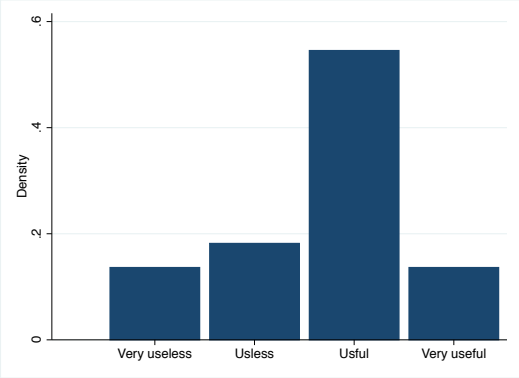


Munz

(b) How useful is the product for neo-Nazis? (Wilcoxon signed-rank test, p= 1.000, N=21)

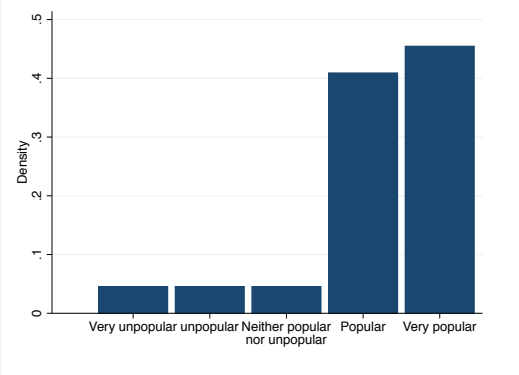


Camille Bloch

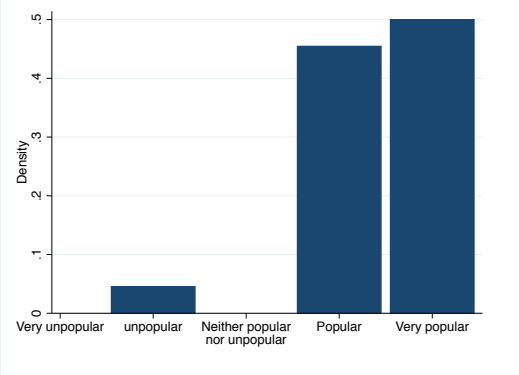


Munz

(c) How popular is the product among students? (Wilcoxon signed-rank test, p= 0.099, N=21)

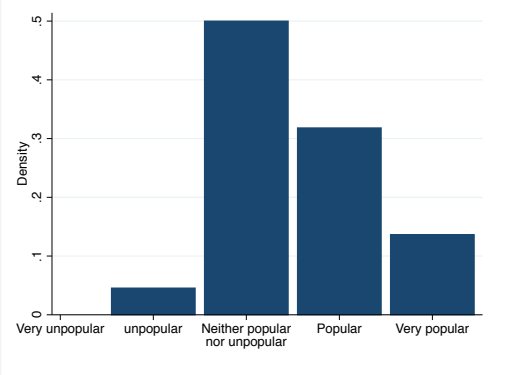


Camille Bloch

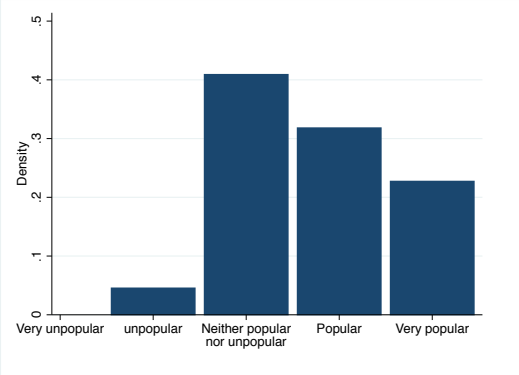


Munz

(d) How popular is the product among neo-Nazis? (Wilcoxon signed-rank test, p=0.293, N=21)

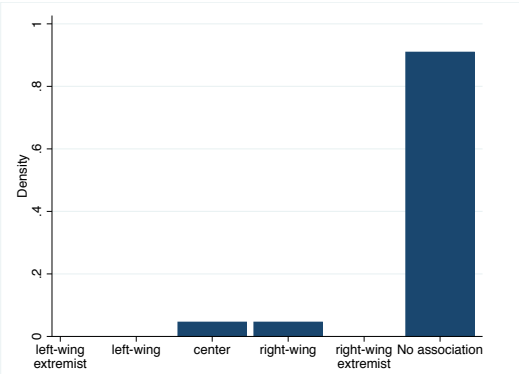


Camille Bloch

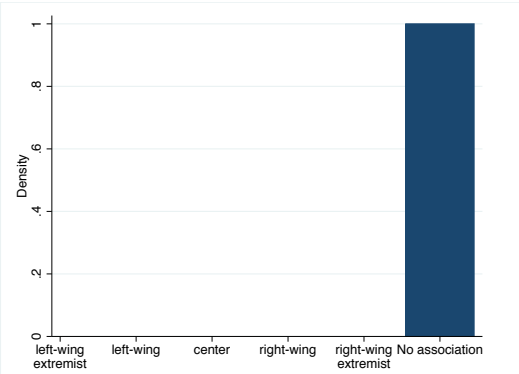


Munz

(e) With which political position would you associate a person that consumes the product? (Wilcoxon signed-rank test, p=0.157, N=21)



Camille Bloch



Munz

Notes: Answers to an online survey with 21 participants. Subjects are drawn from the same subjects pool that is used for Study 2. In figures (e), "No association" was labeled as "I would not associate the product with a specific political position" in the survey.

Table F1: Treatment effects, robustness

Dependent variable:	Pr(Conform to target)			WTP for target's product		
	(1)	(2)	(3)	(4)	(5)	(6)
1 = undesirable treatment	-0.147*** (-2.66)	-0.155*** (-2.81)	-0.191*** (-3.47)	-0.236 (1.59)	-0.265* (-1.86)	-0.323** (-2.24)
Constant	0.624*** (16.09)			0.085 (0.83)		
Log(sigma)				1.320*** (15.17)	1.294*** (15.13)	1.279*** (14.96)
N	308	308	308	308	308	308
Session FE	No	Yes	Yes	No	Yes	Yes

*Notes: (1)-(3): Linear regressions of probability to choose the same product as the target when none of the two products are connected with any payment on a treatment dummy. These specifications include the 7 observations in which subjects made choices that are not monotone in money. (4)-(6): Tobit regressions (left-censored at -3CHF, n=17; right-censored at +3CHF, n=10) of willingness to pay to receive the same product as the target instead of the other product (WTP for target's product) on a treatment dummy. These specifications include the 7 observations in which subjects made choices that are not monotone in money. Given that these observations have multiple switching points, it is unclear how to construct the WTP. I take the average of the WTPs calculated based on the first and on the last switching point. t-statistics in parentheses; standard errors are clustered at subject level (168 clusters); * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Table F2: Treatment effects for intelligence and moral values, robustness

Dependent variable:	Pr(Conform to target)			WTP for target's product		
	(1)	(2)	(3)	(4)	(5)	(6)
1 = undesirable treatment	-0.244*** (-3.11)	-0.252*** (-3.24)	-0.262*** (-3.42)	-0.592*** (-2.71)	-0.610*** (-2.89)	-0.634*** (-3.05)
1 = intelligence round	-0.094 (-1.21)	-0.094 (-1.21)	-0.004 (-0.05)	-0.213 (-1.00)	-0.201 (-0.95)	-0.021 (-0.09)
1 = undesirable * 1= intelligence	0.194* (1.69)	0.194* (1.68)	0.142 (1.24)	0.711** (2.32)	0.688** (2.26)	0.626** (2.04)
Log(sigma)				1.307*** (15.39)	1.281*** (15.32)	1.263*** (15.04)
p-value (undesirable) + (undesirable*intelligence) = 0	0.531	0.473	0.150	0.565	0.703	0.969
N	308	308	308	308	308	308
Session FE	No	Yes	Yes	No	Yes	Yes
Target choice controls	No	No	Yes	No	No	Yes

*Notes: (1)-(3): Linear regressions of probability to choose the same product as the target when none of the two products are connected with any payment on a treatment dummy. These specifications include the 7 observations in which subjects made choices that are not monotone in money. (4)-(6): Tobit regressions (left-censored at -3CHF, n=17; right-censored at +3CHF, n=10) of willingness to pay to receive the same product as the target instead of the other product on a treatment dummy. These specifications include the 7 observations in which subjects made choices that are not monotone in money. Given that these observations have multiple switching points, it is unclear how to construct the WTP. I take the average of the WTPs calculated based on the first and on the last switching point. t-statistics in parentheses; standard errors are clustered at subject level (168 clusters); * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Table F3: Treatment effects, robustness checks

Dependent variable:	**Pr(Munz)**		**WTP Munz**	
	(1)	(2)	(3)	(4)
1 = undesirable treatment	-0.129** (-1.98)	-0.133** (-2.09)	-0.315** (-2.08)	-0.326** (-2.26)
Constant	0.566*** (11.87)		0.077 (0.70)	
Log(sigma)			1.154*** (20.42)	1.093*** (20.43)
N	232	232	232	232
Session Fixed Effects	No	Yes	No	Yes

*Notes: (1) and (2): Linear regressions of probability to choose Munz chocolate when none of the two products are connected with any payment on a treatment dummy. These specifications include the 6 subjects that made choices that are not monotone in money. Robust standard errors are used. (3) and (4): Tobit regressions (left-censored at CHF -3, n=5; right-censored at CHF +3, n=9) of willingness to pay to receive the Munz chocolate instead of the Camille Bloch chocolate (WTP Munz) on a treatment dummy. These specifications include the 6 subjects that made choices that are not monotone in money. Given that these observations have multiple switching points, it is unclear how to construct the WTP. I take the average of the WTPs calculated based on the first and on the last switching point. * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Table F4: Treatment effects and similarity, Study 1

Dependent variable:	**Pr(Conform to target)**			**WTP for target's product**		
	(1)	(2)	(3)	(4)	(5)	(6)
1 = intelligence round	0.104 (0.76)	-0.015 (-0.18)	0.135 (0.86)	0.167 (0.48)	-0.010 (-0.05)	-0.113 (-0.28)
Similarity intelligence * 1 = intelligence	0.022 (0.12)		-0.070 (-0.36)	0.560 (1.22)		0.591 (1.23)
Similarity moral values * 1 = moral values	0.135 (1.10)		0.264** (2.16)	0.452 (1.29)		0.744* (1.97)
1 = undesirable treatment		-0.290*** (-3.76)	-0.334*** (-4.20)		-0.652*** (-3.04)	-0.774*** (-3.29)
1 = undesirable * 1= intelligence		0.161 (1.41)	0.193 (1.64)		0.643** (2.07)	0.833** (2.52)
Log(sigma)				1.290*** (14.96)	1.277*** (14.85)	1.263*** (14.92)
p-value (undesirable) + (undesirable*intelligence) = 0		0.121	0.105		0.966	0.793
N	301	301	301	301	301	301
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
Target choice controls	Yes	Yes	Yes	Yes	Yes	Yes

*Notes: (1)-(3): Linear regressions of probability to choose the same product as the target when none of the two products are connected with any payment on a dummy that captures whether the target was selected due to his intelligence or his moral values (1=intelligence round), similarity for moral values and for intelligence, a treatment dummy (1 = undesirable treatment), and an interaction between this binary variable and the treatment dummy (1=undesirable*1=intelligence). (4)-(6): Tobit regressions (left-censored at CHF -3, n=17; right-censored at CHF +3, n=10) of willingness to pay to receive the same product as the target instead of the other product on the same set of independent variables. Similarity moral values is defined as 1-|donation – target's donation|/6. Similarity intelligence is defined as 1-|intelligence scores – target's intelligence scores|/8. Because similarity moral values is only important for the moral values round, and similarity intelligence is only important for the intelligence round, these measures are interacted with 1-(1=intelligence round) and (1=intelligence round), respectively. t-statistics in parentheses; standard errors are clustered at subject level (168 clusters); * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Table F5: Treatment effects and similarity, Study 2

Dependent variable:	Pr(Munz)					WTP Munz				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Political position	-0.181 (-0.90)				-0.220 (-0.94)	0.447 (0.95)				0.032 (0.06)
Right-wing extremism		0.113 (0.49)			0.091 (0.36)		0.674 (1.24)			0.557 (0.97)
Left-wing extremism			0.088 (0.55)		0.025 (0.14)			-0.382 (-1.07)		-0.368 (-0.94)
Racism				0.086 (0.17)	0.220 (0.45)				1.296 (1.41)	0.894 (0.85)
1 = undesirable treatment	-0.136** (-2.11)	-0.140** (-2.14)	-0.140** (-2.16)	-0.136** (-2.10)	-0.140** (-2.13)	-0.335** (-2.27)	-0.355** (-2.39)	-0.319** (-2.15)	-0.332** (-2.25)	-0.335** (-2.25)
N	226	226	226	226	226	226	226	226	226	226
Session Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

*Notes: (1) - (5): Linear regressions of probability to choose Munz chocolate when none of the two products are connected with any payment on different measures of similarity and a treatment dummy. Robust standard errors are used. (6) - (10): Tobit regressions (left-censored at -3CHF, n=5; right-censored at +3CHF, n=9) of willingness to pay to receive the Munz chocolate instead of the Camille Bloch chocolate on different measures of similarity and a treatment dummy. Political position is in between 0 (=very left-wing) and 1 (=very right-wing), with mean 0.463 (s.d.=0.160). To measure Right-wing (left-wing) extremism, subjects are asked how much they agree (from 1= strongly disagree to 7=strongly agree) with three statements that are typically accepted by right-wing (left-wing) extremists. I took the average of the three answers, and divided it by 7. The mean of right-wing extremism is 0.103 (s.d.=0.139) and the mean of left-wing extremism is 0.331 (s.d.= 0.216). To measure racism, subjects are asked how much they agree with seven statements related to racism, adapted from the Modern Racism Scale (McConahay, 1986). I took the average of all answers, and divided it by 7. The mean of racism is 0.301 (s.d.=0.082). t-statistics in parentheses; * - p < 0.1; ** - p < 0.05; *** - p < 0.01.*

Table F6: Perception of the products, Study 1

(a) IQ and Pro-sociality

	Camille Bloch/Munz						**Cup/USB Stick**					
	Product chosen by target, mean rating			*Product not chosen by target, mean rating*			*Product chosen by target, mean rating*			*Product not chosen by target, mean rating*		
	Desirable treatment	*Undesirable treatment*	*rank-sum z (p)*	*Desirable treatment*	*Undesirable treatment*	*rank-sum z (p)*	*Desirable treatment*	*Undesirable treatment*	*rank-sum z (p)*	*Desirable treatment*	*Undesirable treatment*	*rank-sum z (p)*
What do you think, how much does the product currently cost? (in CHF)	3.94	3.88	0.442 (0.658)	3.88	3.70	1.003 (0.316)	8.71	7.56	1.519 (0.129)	7.56	7.26	0.181 (0.856)
How good is the quality of the raw materials used? (1 = "very low quality"; 5 = "very high quality")	3.71	3.52	1.519 (0.129)	3.64	3.26	2.598 (0.009)	3.17	3.01	1.034 (0.301)	3.23	3.07	.799 (0.425)
How well is the product processed? (1 = "very low quality"; 5 = "very high quality")	3.81	3.67	1.454 (0.146)	3.76	3.45	2.232 (0.026)	3.32	3.22	0.681 (0.496)	3.39	3.15	1.638 (0.101)
When you think about using/eating the product, are you disgusted? (1 = "very disgusted"; 5 = "not disgusted at all")	4.42	4.15	0.974 (0.330)	4.33	4.08	1.032 (0.302)	4.51	4.45	-0.179 (0.858)	4.52	4.41	0.496 (0.620)

Notes: "rank-sum"-columns give the z-value and p-value of a Wilcoxon rank-sum test.

(b) Pro-sociality only

	Camille Bloch/Munz						Cup/USB Stick					
	Product chosen by target, mean rating			*Product not chosen by target, mean rating*			*Product chosen by target, mean rating*			*Product not chosen by target, mean rating*		
	Desirable treatment	*Undesirable treatment*	*rank-sum z (p)*	*Desirable treatment*	*Undesirable treatment*	*rank-sum z (p)*	*Desirable treatment*	*Undesirable treatment*	*rank-sum z (p)*	*Desirable treatment*	*Undesirable treatment*	*rank-sum z (p)*
What do you think, how much does the product currently cost? (in CHF)	3.95	3.95	0.461 (0.645)	3.99	3.73	0.948 (0.343)	8.37	7.35	1.170 (0.242)	7.57	7.31	0.094 (0.925)
How good is the quality of the raw materials used? (1 = "very low quality"; 5 = "very high quality")	3.75	3.58	0.420 (0.674)	3.71	3.42	1.058 (0.290)	3.10	2.93	0.859 (0.390)	3.20	2.98	0.826 (0.409)
How well is the product processed? (1 = "very low quality"; 5 = "very high quality")	3.79	3.62	0.814 (0.416)	3.75	3.42	1.238 (0.216)	3.29	3.13	0.765 (0.444)	3.31	3.13	0.836 (0.403)
When you think about using/eating the product, are you disgusted? (1 = "very disgusted"; 5 = "not disgusted at all")	4.46	4.19	0.041 (0.967)	4.36	3.96	0.646 (0.518)	4.57	4.52	-0.193 (0.847)	4.65	4.46	1.166 (0.244)
Does the producer of the good promote conservative Christian values (e.g, fighting abortions and marriage for same-sex couples)? (1 = "yes, strongly "; 5 = "no, not at all")	3.82	3.96	-0.509 (0.611)	3.57	3.88	-1.054 (0.292)	4.08	4.13	-0.193 (0.847)	4.16	4.02	0.757 (0.449)
How popular is the product among conservative Christians? (1 = "not popular at all"; 5 = "very popular")	3.29	3.04	1.069 (0.285)	3.36	3.19	0.822 (0.411)	3.18	3.24	-0.600 (0.549)	3.22	3.37	-0.802 (0.423)
How popular is the product among Atheists? (1 = "not popular at all"; 5 = "very popular")	3.43	3.15	1.239 (0.215)	3.39	3.12	1.528 (0.127)	3.35	3.5	-0.859 (0.39)	3.29	3.37	-0.528 (0.597)

Notes: "rank-sum"-columns give the z-value and p-value of a Wilcoxon rank-sum test.

(c) IQ only

	Camille Bloch/Munz						Cup/USB Stick					
	Product chosen by target, mean rating			*Product not chosen by target, mean rating*			*Product chosen by target, mean rating*			*Product not chosen by target, mean rating*		
	Desirable treatment	*Undesirable treatment*	*rank-sum z (p)*	*Desirable treatment*	*Undesirable treatment*	*rank-sum z (p)*	*Desirable treatment*	*Undesirable treatment*	*rank-sum z (p)*	*Desirable treatment*	*Undesirable treatment*	*rank-sum z (p)*
What do you think, how much does the product currently cost? (in CHF)	3.93	3.84	0.233 (0.816)	3.81	3.68	0.552 (0.581)	9.29	7.92	0.963 (0.336)	7.54	7.18	0.178 (0.859)
How good is the quality of the raw materials used? (1 = "very low quality"; 5 = "very high quality")	3.68	3.49	1.528 (0.126)	3.60	3.17	2.310 (0.021)	3.29	3.15	0.685 (0.493)	3.29	3.22	0.280 (0.780)
How well is the product processed? (1 = "very low quality"; 5 = "very high quality")	3.82	3.70	1.215 (0.224)	3.76	3.47	1.847 (0.065)	3.39	3.37	0.203 (0.839)	3.54	3.19	1.661 (0.097)
When you think about using/eating the product, are you disgusted? (1 = "very disgusted"; 5 = "not disgusted at all")	4.40	4.13	1.139 (0.255)	4.32	4.15	0.804 (0.421)	4.39	4.33	-0.062 (0.951)	4.29	4.33	-0.561 (0.575)
How popular is the product among intelligent people? (1 = "not popular at all"; 5 = "very popular")	3.16	3.19	-0.025 (0.980)	3.22	3.09	1.165 (0.244)	3.75	3.30	1.936 (0.053)	3.68	3.59	0.314 (0.754)
How popular is the product among stupid people? (1 = "not popular at all"; 5 = "very popular")	3.54	3.13	2.542 (0.011)	3.42	3.04	2.358 (0.018)	3.50	3.22	1.463 (0.143)	3.32	3.48	-0.688 (0.491)

Notes: "rank-sum"-columns give the z-value and p-value of a Wilcoxon rank-sum test.

Table F7: Perception of the products, Study 2

	Camille Bloch			Munz		
	Neutral treatment, mean rating	*Immoral treatment, mean rating*	*rank-sum, z (p-value)*	*Neutral treatment, mean rating*	*Immoral treatment, mean rating*	*rank-sum, z (p-value)*
"What do you think, how much does a pack of Camille Bloch Torino (Munz) chocolate bars currently cost at Migros?" (in CHF)	3.88	4.01	-0.559 (0.576)	3.77	3.9	-0.757 (0.449)
"How healthy are the products?" (1 = "very unhealthy"; 5 = "very healthy")	1.94	1.94	-0.027 (0.978)	1.9	1.95	-0.518 (0.604)
„How long can you store them?" (1 = " spoils soon "; 5 = "long storage life")	4.39	4.26	1.333 (0.182)	4.42	4.23	1.997 (0.046)
„What is the quality of the raw materials used?" (1 = "very low quality"; 5 = "very high quality")	3.5	3.47	0.166 (0.868)	3.41	3.46	-0.389 (0.697)
„How sustainable are the raw materials used?" (1 = "very sustainable"; 5 = "not sustainable at all")	3.38	3.22	1.430 (0.153)	3.41	3.22	1.670 (0.095)
"What is the quality of the processing?" (1 = "very low quality"; 5 = "very high quality")	3.68	3.78	-0.872 (0.383)	3.55	3.64	-1.159 (0.247)
„When you think about eating the Camille Bloch (Munz) chocolate, are you digusted?" (1 = "very disgusted"; 5 = "not disgusted at all")	4.34	4.44	-0.693 (0.488)	4.46	4.6	-0.803 (0.422)

See next page for the rest of the table.

	Camille Bloch			**Munz**		
	Neutral treatment, mean rating	*Immoral treatment, mean rating*	*rank-sum, z (p-value)*	*Neutral treatment, mean rating*	*Immoral treatment, mean rating*	*rank-sum, z (p-value)*
"If you were to eat the chocolate in public, would people associate you with right-wing extremism?" (1 = "Yes, for sure"; 5 = "No, certainly not")	4.93	4.86	-0.024 (0.981)	4.89	4.82	0.625 (0.532)
"If you were to eat the chocolate in front of your family or friends, would you associate it with right-wing extremism?" (1 = "Yes, for sure"; 5 = "No, certainly not")	4.95	4.91	0.276 (0.782)	4.93	4.89	-0.031 (0.975)
"What do you think, does Camille Bloch (Munz) in any way promote right-wing extremism (for example, through party donations, employment of right-wing extremists, public support for right-wing extremist concerns)?" (1 = "Yes, very strong"; 5 = "No, not at all")	4.39	4.45	-0.547 (0.585)	4.39	4.34	0.254 (0.800)
"What do you think, does Camille Bloch (Munz) in any way discriminate against minorities (for example, in the recruitment, pay and promotion of employees)?" (1 = "Yes, very strong"; 5 = "No, not at all")	3.96	4.12	-1.445 (0.149)	3.97	4.04	-0.831 (0.406)

Notes: "rank-sum"-columns give the z-value and p-value of a Wilcoxon rank-sum test.

Appendix G – Proofs Chapter 3

Proposition. The relationship between $x_g^{A\cup B}$ and $x_b^{A\cup B}$ is characterized by a threshold $\overline{\Delta u}$ such that $x_g^{A\cup B} > \max(x_b^{A\cup B})$ if $\Delta u < \overline{\Delta u}$ and $x_g^{A\cup B} = \max(x_b^{A\cup B}) = \frac{\delta}{1-\gamma}$ if $\Delta u \geq \overline{\Delta u}$. Furthermore, $\overline{\Delta u}$ is increasing in α.

Proof: I will first derive the equilibria for the case that the A_c-types have undesirable characteristics, *c=b*. Note that $x_b^{A_c} = 1$ and the restriction to monotonic equilibria implies that $\rho(B) = 0$ and therefore $x_b^B = 0$. The A-types choose A iff $\Delta u \geq \rho(A)\alpha(v^{-c} - v^b)$ with $\rho(A) = \frac{\gamma}{\delta x^A + \gamma}$. There are three potential equilibria:

- $x_b^A = 0$. In this case, $\rho(A) = 1$. This is an equilibrium iff $\Delta u \leq \alpha(v^{-c} - v^b)$.
- $x_b^A = 1$. In this case, $\rho(A) = \frac{\gamma}{\delta+\gamma}$. This is an equilibrium iff $\Delta u \geq \frac{\gamma}{\delta+\gamma}\alpha(v^{-c} - v^b)$.
- $x_b^A \in (0,1)$. In this case, $\rho(A) = \frac{\gamma}{\delta x_b^A + \gamma}$ and $x_b^A = \frac{\gamma}{\delta}\left(\frac{\alpha}{\Delta u}(v^{-c} - v^b) - 1\right)$. This is an equilibrium iff $\frac{\gamma}{\delta+\gamma}\alpha(v^{-c} - v^b) < \Delta u < \alpha(v^{-c} - v^b)$.

Note that for each value of Δu there is at least one monotone equilibrium, and at most three equilibria.[125] The share of A- and B-types that choose product A then is given by:

$$x_b^{A\cup B} = \begin{cases} 0 & if\ \Delta u < \frac{\gamma}{\delta+\gamma}\alpha(v^{-c} - v^b) \\ \left\{0, \frac{\gamma}{1-\gamma}\left(\frac{\alpha}{\Delta u}(v^{-c} - v^b) - 1\right), \frac{\delta}{1-\gamma}\right\} & if\ \Delta u \in \left[\frac{\gamma}{\delta+\gamma}\alpha(v^{-c} - v^b), \alpha(v^{-c} - v^b)\right] \\ \frac{\delta}{1-\gamma} & if\ \Delta u > \alpha(v^{-c} - v^b) \end{cases}$$

with $\frac{\gamma}{1-\gamma}\left(\frac{\alpha}{\Delta u}(v^{-c} - v^b) - 1\right) \leq \frac{\delta}{1-\gamma}$ for $\Delta u \geq \frac{\gamma}{\delta+\gamma}\alpha(v^{-c} - v^b)$. Therefore, $\max(x_b^{A\cup B})$ is 0 for $\Delta u < \frac{\gamma}{\delta+\gamma}\alpha(v^{-c} - v^b)$ and $\frac{\delta}{1-\gamma}$ for $\Delta u \geq \frac{\gamma}{\delta+\gamma}\alpha(v^{-c} - v^b)$. Figure G1 illustrates how $x_b^{A\cup B}$ depends on Δu.

[125] The number of equilibria could be reduced by restricting attention to undominated equilibria, as Bénabou & Tirole (2011) do. This equilibrium refinement criterion eliminates equilibria that are Pareto-dominated (weakly lower payoffs for all types, and a strictly lower payoff for at least one of them). The equilibria would then be given by: $x_u^A = 0$ for $\Delta u < \alpha(v^{-c} - v^u)$, $x_u^A \in \{0,1\}$ for $\Delta u = \alpha(v^{-c} - v^u)$ and $x_u^A = 1$ for $\Delta u > \alpha(v^{-c} - v^u)$. If in addition the A_c-types are assumed to care about their image (such that $x_u^{A_c} = 1$ still holds), the equilibrium $x_u^A = 0$ for $\Delta u = \alpha(v^{-c} - v^u)$ would be Pareto-dominated by $x_u^A = 1$, and the equilibrium would be unique. However, none of this is necessary to proof the Proposition.

Next, I will derive the equilibria for the case that the A_c-types have desirable characteristics, c=g. Note that $x_g^{A_c} = 1$ and the restriction to monotonic equilibria implies that $\rho(B) = 0$ and therefore $x_g^A = 1$. The B-types choose B iff $\Delta u \geq \rho(A)\alpha(v^g - v^{-c})$ with $\rho(A) = \frac{\gamma}{1-(1-\gamma-\delta)(1-x_g^B)}$. There are three potential equilibria:

- $x_g^B = 0$. In this case, $\rho(A) = \frac{\gamma}{\gamma+\delta}$. This is an equilibrium iff $\Delta u \geq \frac{\gamma}{\gamma+\delta}\alpha(v^g - v^{-c})$
- $x_g^B = 1$. In this case, $\rho(A) = \gamma$. This is an equilibrium iff $\Delta u \leq \gamma\alpha(v^g - v^{-c})$.
- $x_g^B \in (0,1)$. In this case, $\rho(A) = \frac{\gamma}{1-(1-\gamma-\delta)(1-x_g^B)}$ and $x_g^B = \frac{\gamma\alpha(v^g - v^{-c}) - \Delta u(\gamma+\delta)}{\Delta u(1-\gamma-\delta)}$. This is an equilibrium iff $\gamma\alpha(v^g - v^{-c}) < \Delta u < \frac{\gamma}{\gamma+\delta}\alpha(v^g - v^{-c})$.

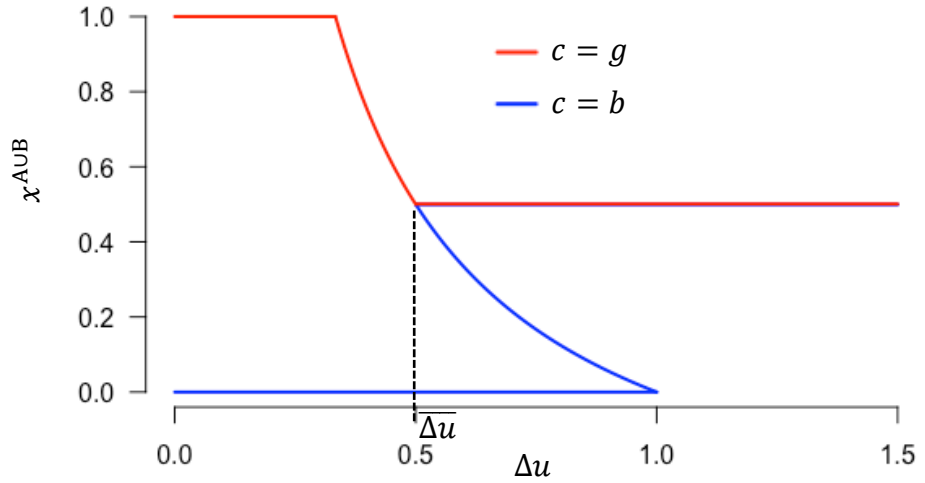
Note that for each value of Δu there is a unique monotone equilibrium. The number of A- and B-types that choose Product A then is given by:

$$x_g^{A\cup B} = \begin{cases} 1 & if\ \Delta u \leq \gamma\alpha(v^g - v^{-c}) \\ \frac{\gamma}{1-\gamma}\left(\frac{\alpha}{\Delta u}(v^g - v^{-c}) - 1\right) & if\ \Delta u \in \left(\gamma\alpha(v^g - v^{-c}), \frac{\gamma}{\gamma+\delta}\alpha(v^g - v^{-c})\right) \\ \frac{\delta}{1-\gamma} & if\ \Delta u \geq \frac{\gamma}{\gamma+\delta}\alpha(v^g - v^{-c}) \end{cases}$$

with $\frac{\gamma}{1-\gamma}\left(\frac{\alpha}{\Delta u}(v^g - v^{-c}) - 1\right) > \frac{\delta}{1-\gamma}$ for $\Delta u < \frac{\gamma}{\gamma+\delta}\alpha(v^g - v^{-c})$. Figure G1 illustrates how $x_g^{A\cup B}$ depends on Δu.

To finish the proof, define $\overline{\Delta u} = \alpha\frac{\gamma}{\delta+\gamma}\max(v^{-c} - v^b, v^g - v^{-c})$ and note that $x_g^{A\cup B} > \max(x_b^{A\cup B})$ for $\Delta u < \overline{\Delta u}$, that $x_g^{A\cup B} = \max(x_b^{A\cup B}) = \frac{\delta}{1-\gamma}$ for $\Delta u \geq \overline{\Delta u}$ and that $\overline{\Delta u}$ is increasing in α.

Figure G1



Note: $x^A + x^B$ for c=g and for c=b for parameters $\gamma = \delta = 1/3$, $v^g - v^{-c} = v^{-c} - v^b = \alpha = 1$

Appendix H – Additional figures Chapter 3

Figure H1: Observers' instructions

Remember that in Part 1, all participants did a test to measure their intelligence.

The participant with the highest intelligence score and another, randomly drawn participant made a choice between two products, Camille Bloch chocolate and Munz chocolate. You will observe the product choice and portrait picture of one of these two participants. Which participant you observe is determined by a (virtual) coin flip:

- With a probability of 50% heads comes up, and you observe the choice and the picture of the participant with the highest intelligence score,
- With a probability of 50% tails comes up, and you observe the choice and the picture of the other randomly selected participant.

The computer will *not* tell you whether heads or tails came up. *However, you will learn which product was chosen by the participant with the highest intelligence score.*

Example: Suppose that the participant with the lowest intelligence score chose the Camille Bloch chocolate, the other randomly selected participant chose **Munz chocolate**. If tails comes up, you would see the following screen:

The participant with the **lowest intelligence** score chose **the Camille Bloch chocolate**.

The following participant is drawn:

The person on the picture choose **the Munz chocolate**.

Figure H2: Presentation of product choices to neo-Nazis

Left option (unattractive symbols for neo-Nazis)	*Right option (attractive symbols for neo-Nazis)*	*Share Right*
Intenso Rainbow Line 4GB USB Stick Blau (Art. Nr. 3502450) 	Butlers HENKELBECHER KREUZ Grau (Art Nr. 10210598) 	0.1
Intenso Rainbow Line 4GB USB Stick Blau (Art. Nr. 3502450) 	Kahla Tasse 0,18l Rot-Weiss (Art Nr. 27508877) 	0.6
Kahla Colore Tasse 0,25l Grau (Art Nr. 204708A70705C) 	Butlers HENKELBECHER KREUZ Grau (Art Nr. 10210598) 	0.2
Kahla Colore Tasse 0,25l Schokobraun (Art Nr. 204708A72605C) 	Kahla Tasse 0,18l Rot-Weiss (Art Nr. 27508877) 	0.7

See next page for the rest of the table.

Left option *(unattractive symbols for neo-Nazis)*	*Right option* *(attractive symbols for neo-Nazis)*	*Share Right*
Intenso Rainbow Line 8GB USB Stick Grün (Art. Nr. 3502460) 	Butlers HENKELBECHER KREUZ Grau (Art Nr. 10210598) 	0.0
Kahla Colore Tasse 0,25l Schokobraun (Art Nr. 204708A72605C) 	Butlers HENKELBECHER KREUZ Grau (Art Nr. 10210598) 	0.3
Intenso Rainbow Line 8GB USB Stick Grün (Art. Nr. 3502460) 	Intenso Rainbow Line 4GB USB Stick Blau (Art. Nr. 3502450) 	0.2

Notes: The colors show the symbols that might make a product more (red) or less (blue) attractive for neo-Nazis. "Rainbow" and "Schokobraun" (=chocolate brown) are symbols of cultural diversity (however, brown is also the color of the Nazi Party, besides black-white-red). "HENKELBECHER KREUZ" sounds similar as "Hakenkreuz" (=swastika), "Rot-Weiss" (=red-white) are the colors of the Swiss flag, and 88 is a well known Nazi symbol for Heil Hitler (H is the eight letter in the Alphabet). The sample of right-wing extremists did not like the "HENKELBECHER KREUZ." The last two choices differ only little in terms of symbols. The left option dominates the right option, if one abstracts from differences in colors. I added this option for potential future research on the potential limits of identity signaling. (Does disconformity occur when neo-Nazis adopt a product that dominates another product?)

Figure H3: Targets' choices are revealed

As announced, 10 participants from a previous study had to chose between two product.

These participants are **neo-Nazis**. The neo-Nazis were recruited on the right-wing extremist websites [redacted] and [redacted]. The neo-Nazis chose between products A and B. The chosen product was delivered to the participants while maintaining their anonymity.

[Neutral treatment: These participants were recruited on the internet and chose between products A and B. The chosen product was delivered to the participants.]

- 70 percent of participants are men.
- The average age is 33 years.
- Eighty percent of participants have completed at most upper secondary education.

The 10 participants chose between the following products:

Pack Camille Bloch Torino
5 x 23g (Art. Nr. 3000652)

Pack Munz Praliné-Prügeli Milch
5 x 23g (Art. Nr. 3928213)

Result: **9 out of 10 neo-Nazis** [Neutral treatment: **participants**] chose **Munz Praliné-Prügeli Milch**.

Now, **Part 3** follows: You have to chose among these two product. Which product do you want?

○ Pack Camille Bloch Torino
○ Pack Munz Praliné-Prügeli Milch

Notes: Translated from German.

Appendix I – Additional results Chapter 4

Figure I1 – Distribution of Player A choices in Experiment 1

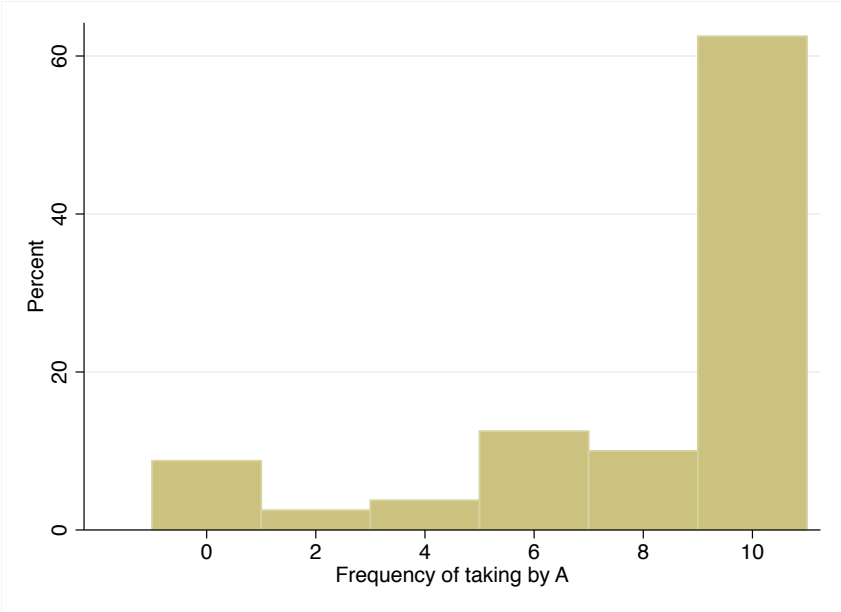


Figure I2 – Distribution of Player B choices in Experiment 2

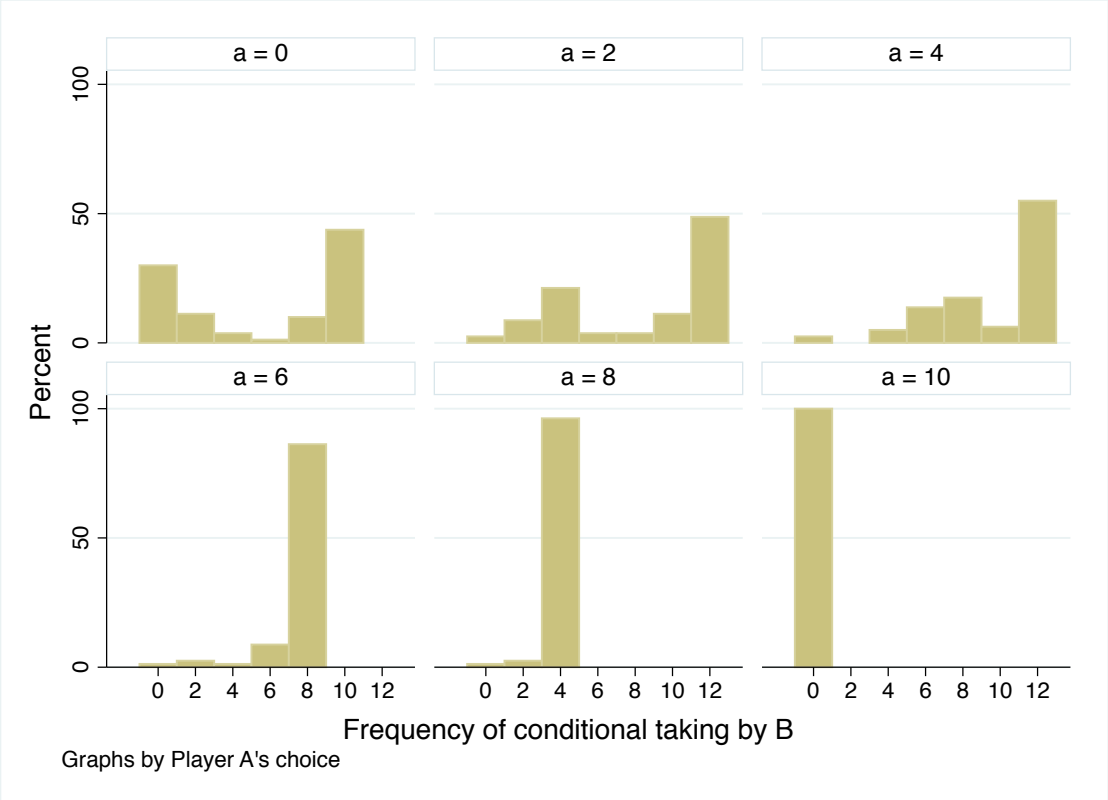
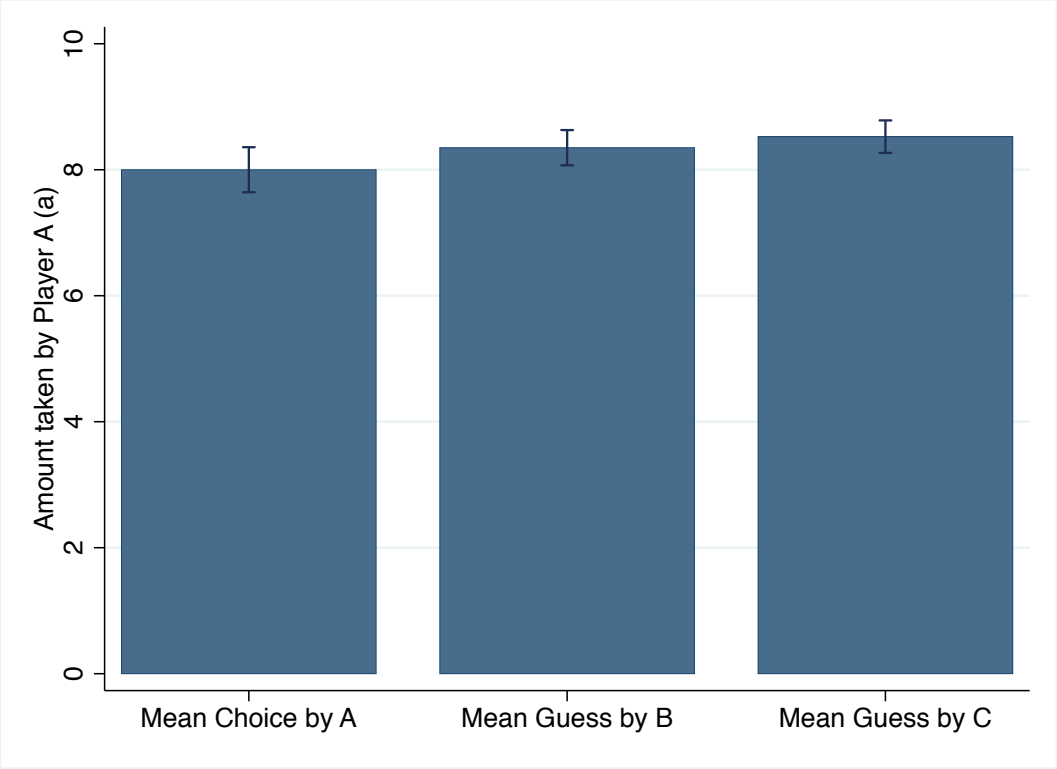


Figure I3 – Taking by Player A and corresponding beliefs of Players B and C



Appendix J – Proofs Chapter 4

Proposition 1: For any $(x, \hat{p})$ *in* $argmax_{x\in[0,K],\hat{p}\in[0,1]}$ U$(x, \hat{p})$, $\hat{p} \leq p$.

Proof: For any $\hat{p} \in (p, 1]$ and any $x \in [0, K]$ we have that $U(x, p) > U(x, \hat{p})$:

$$U(x,p) - U(x,\hat{p}) = \left(E_p(d) - E_{\hat{p}}(d)\right) v(1-x)$$
$$+\big(C(\hat{p}-p) - C(0)\big) + (P(0) - P(\hat{p}-p))$$

First, note that $\left(E_p(d) - E_{\hat{p}}(d)\right) v(1-x) > 0$ as $E_p(d) > E_{\hat{p}}(d)$ and $v(.) > 0$. Second, we have $C(\hat{p}-p) - C(0) > 0$ as $C'(.) > 0$. Third, $P'(.) < 0$ implies that $P(0) - P(\hat{p}-p) > 0$. Therefore $U(x,p) - U(x,\hat{p}) > 0$, and $\hat{p} \in (p, 1]$ cannot be a solution to (1).

Proposition 2: Take K, K' *in* $(0,1]$ *with* $K' < K$ *and suppose that there is a unique solution to (1) for both* K *and* K'*, then* $\hat{p}' \leq \hat{p}$.

Proof: Define $(x, \hat{p}) = argmax_{x\in[0,K],\hat{p}\in[0,1]}$ U$(x, \hat{p})$ and $(x', \hat{p}') = argmax_{x\in[0,K\prime],\hat{p}\in[0,1]}$ U$(x, \hat{p})$.

Case i) Suppose $x \leq K'$, then $(x, \hat{p})$ is the solution of $max_{x\in[0,K\prime],\hat{p}\in[0,1]}\, U(x, \hat{p})$, so $\hat{p}' = \hat{p}$.

Case ii) Suppose $x(K) > K'$. Then $x(K') < x(K)$ (and $1 - x(K') > 1 - x(K)$). Note that the relevant Karush–Kuhn–Tucker conditions for the problem are (Proposition 1 implies that the condition $\hat{p} \leq 1$ is not binding):

$$-\mathrm{P}'(p-\hat{p}) - (d_H - d_L)v(1-x) + C'(p-\hat{p}) + \lambda = 0 \text{ (I)}$$
$$\lambda\hat{p} = 0 \text{ (II)}$$
$$\hat{p} \geq 0 \text{ (III)}$$
$$\lambda \geq 0 \text{ (IV)}$$

Case iia) Suppose $\hat{p}' = 0$, then $\hat{p}' \leq \hat{p}$ due to condition (III).

Case iib) Suppose $\hat{p} = 0$. Then (I) implies

$$(d_H - d_L)v(1-x) \geq C'(p) - \mathrm{P}'(p) \text{ (I')}$$

Note that $(d_H - d_L)v(1-x') > (d_H - d_L)v(1-x)$ as $v'(.) > 0$. This, together with (I') and the assumptions on C and P imply $(d_H - d_L)v(1-x') > C'(p) - \mathrm{P}'(p) \geq C'(p-\hat{p}') - \mathrm{P}'(p-\hat{p}')$. Then by (I), $\lambda' > 0$. Then by (II) $\hat{p}' = 0$.

Case iic) Suppose $\hat{p}, \hat{p}' > 0$. Then by (II) $\lambda = \lambda' = 0$. Then (I) simplifies to:

$$(d_H - d_L)v(1-x) = C'(p-\hat{p}) - \mathrm{P}'(p-\hat{p}) \text{ (I'')}$$
$$(d_H - d_L)v(1-x') = C'(p-\hat{p}') - \mathrm{P}'(p-\hat{p}') \text{ (I''')}$$

Combining (I''), (I''') and $(d_H - d_L)v(1-x') > (d_H - d_L)v(1-x)$ implies:

$$C'(p-\hat{p}') - \mathrm{P}'(p-\hat{p}') > C'(p-\hat{p}') - \mathrm{P}'(p-\hat{p}) \text{ (V)}$$

(V) together with $P'(.) < 0$, $P''(.) \leq 0$, $C'(.) > 0$ and $C''(.) > 0$ implies $\hat{p}' \leq \hat{p}$.

Proposition 3: $\hat{p}^N = p$ *is the unique solution to* $\max_{\hat{p}^N\in[0,1]} U(\hat{p}^N)$.

Proof: For any $\hat{p}^N \in (p, 1]$ we have that $U(p) - U(\hat{p}^N) = P(0) - P(\hat{p}^N - p) + C(\hat{p}^N - p) - C(0) > 0$ due to $C'(.) > 0$ and $P'(.) < 0$.

For any $\hat{p}^N \in [0, p)$ we have that $U(p) - U(\hat{p}^N) = P(0) - P(p - \hat{p}^N) + C(p - \hat{p}^N) - C(0) > 0$ due to $C'(.) > 0$ and $P'(.) < 0$.

Curriculum Vitae

Personal details

Florian Schneider

Date of birth: 26.07.1986

Education

09/2014–04/2020	Doctoral program at the Zurich Graduate School of Economics, University of Zurich
01/2019-06/2019	Invited research visit at Rady School of Management, University of California San Diego
02/2012 - 07/2014	Master of Science in Business and Economics, University of Basel
09/2009 - 08/2011	Bachelor of Science in Business Administration, Zurich University of Applied Sciences

Professional experience

03/2013 - 07/2014	Research assistant, Prof. Daniel Chen, ETH, Center for Law and Economics, Zurich
09/2011 - 04/2013	Research assistant, Arbeitsmarktbeobachtung Ostschweiz (AMOSA), Zurich